Evaluation of pesticide toxicity: a hierarchical QSAR approach to model the acute aquatic toxicity and avian oral toxicity of pesticides

Thesis

EVALUATION OF PESTICIDE TOXICITY

A hierarchical QSAR approach to model the acute aquatic toxicity and avian oral toxicity of pesticides

Thesis submitted by

Paolo Mazzatorta

Mario Negri Institute for Pharmacological Research, Milano, Italy
in collaboration with the Open University, London, UK

for the degree of

Doctor of Philosophy

under the supervision of

Director of Studies: Dr Emilio Benfenati
Second Supervisor: Dr Mark T.D. Cronin

MAY 2005

Alla mia famiglia,

Dixeris egregie, notum si callida verbum

reddiderit iunctura novum.

- Q. Horatius Flaccus, De arte poetica liber, 47-48 -

SE12
(December 2002)

THE OPEN UNIVERSITY

RESEARCH SCHOOL

Research Degrees in Sponsoring Establishments

Library Authorisation

Part One: Candidate Details

Name: ...Paolo Mazzatorta..... PI: ...W2699048.....

Degree: ...PhD..... Sponsoring Establishment: Istituto di Ricerche Farmacologiche Mario Negri

Thesis title: Evaluation of Pesticide Toxicy: A Hierarchical QSAR Approach To Model Acute Aquatic Toxicity and Oral Avian Toxicity of Pesticides.....

Part Two: Open University Library Authorisation

I confirm that I am willing for my thesis to be made available to readers by the Open University Library, and that it may be photocopied, subject to the discretion of the Librarian.

Signed: Paolo Mazzatorta..... Date: ...2/2/2006.....

Part Three: British Library Authorisation [PhD candidates only]

If you want a copy of your PhD thesis to be available on loan to the British Library Thesis Service as and when it is requested, you must sign a British Library Doctoral Thesis Agreement Form. Please return it to the Research School with this form. The British Library will publicise the details of your thesis and may request a copy on loan from the University Library. Information on the presentation of the thesis is given in the Agreement Form.

Please note the British Library have requested that theses should be printed on one side only to enable them to produce a clear microfilm. The Open University Library sends the fully bound copy of theses to the British Library.

The University has agreed that your participation in the British Library Thesis Service should be voluntary. Please tick either (a) or (b) to indicate your intentions.

☒ I am willing for the Open University to loan the British Library a copy of my thesis. A signed Agreement Form is attached

☐ I do not wish the Open University to loan the British Library a copy of my thesis.

Signed: Paolo Mazzatorta..... Date: ...2/2/2006.....

Mazzatorta, Paolo (2005)
Evaluation of Pesticide Toxicity: A hierarchical QSAR approach to model the acute aquatic toxicity and avian oral toxicity of pesticides.
Mario Negri Institute for Pharmacological Research
via Eritrea 62, 20157 Milano, Italy

ABSTRACT

The thesis aimed to extract information relevant to the hazard and risk assessment of pesticides. In particular, quantitative structure-activity relationship (QSAR) approaches have been used to build up a mathematical model able to predict the aquatic acute toxicity, LC_{50}, and the avian oral toxicity, LD_{50}, for pesticides. Ecotoxicological values were collected from several databases, and screened according to quality criteria.
A hierarchical QSAR approach was applied for the prediction of acute aquatic toxicity. Chemical structures were encoded into molecular descriptors by an automated, seamless procedure available within the OpenMolGRID system. Different linear and non-linear regression techniques were used to obtain reliable and thoroughly validated QSARs. The final model was developed by a counter-propagation neural network coupled with genetic algorithms for variable selection. The proposed QSAR is consistent with McFarland's principle for biological activity and makes use of seven molecular descriptors. The model was assessed thoroughly in test (R^2 = 0.8) and validation sets (R^2 = 0.72), the y-scrambling test and a sensitivity/stability test.
The second endpoint considered in this thesis was avian oral toxicity. As previously, the chemical description of chemicals was generated automatically by the OpenMolGRID system. The best classification model was chosen on the basis of the performances on a validation set of 19 data points, and was obtained from a support vector machine using 94 data points and nine variables selected by genetic algorithms (Error $Rate_{training}$ = 0.021, Error $Rate_{validation}$ = 0.158). The model allowed for a mechanistic estimation of the toxicological action. In fact, several descriptors selected for the final classification model encode for the interaction of the pesticides with other molecules. The presence of hetero-atoms, e.g. sulphur atoms, is correlated with the toxicity, and the pool of descriptor selected is generally dependent from the 3D conformation of the structures. These suggest that, in the case of avian oral toxicity, pesticides probably exert their toxic action through the interaction with some macromolecule and/or protein of the biological system.

Keywords: QSAR, Pesticides, Acute aquatic toxicity, LC_{50}, Rainbow trout, Avian oral toxicity, LD_{50}, Bobwhite quail, Regression, Classification, Neural network, Support vector machine.

ACKNOWLEDGEMENTS

First I would like to express my gratitude to my supervisor at Mario Negri Institute, Emilio Benfenati, who believed in me, and gave me the opportunity to start working at the Institute. I should also like to thank him for all the help and advice in the development of my work. I feel very much in debt to him.

I should also like to thank my supervisor at the John Moores University, Mark Cronin, for his very valuable discussions and comments. He also made an important contribution in reviewing the drafts.

My colleagues and friends at the Mario Negri Institute have always been a source of inspiration, and I wish to thank especially Elena Lo Piparo and Martin Smiesko, for patiently listening to my ideas and for their discerning comments.

The laboratory of chemometrics at the National Institute of Chemistry in Ljubljana is thanked for the development of the original FORTRAN code implementing counter-propagation neural network. In particular, Marjan Vracko is thanked for guiding me in -for me- new fields of chemometrics.

The OpenMolGRID consortium is acknowledged for the development and use of the OpenMolGRID system.

I would also like to thank the partners of the DEMETRA project for the interesting and fruitful collaboration.

Financial support from the European Community, through IMAGETox, OpenMolGRID and DEMETRA projects, is gratefully acknowledged.

Last but not least my family, Elena (again), and my friends for their continuous encouragement, support and advice.

PREFACE

The aim of this thesis was to extract information relevant to the hazard and risk assessment of pesticides. In the introductory chapter some basic information and background knowledge, i.e. the state of the art and from where this study started are given, together with an overview of the rationale and strategy adopted for this research.

The remainder of this thesis describes the principal concepts of project, and presents the major findings obtained during the PhD study. In particular, quantitative structure-activity relationship approaches have been used to build up a mathematical model able to predict the aquatic acute toxicity, i.e. LC_{50}, and the avian oral toxicity, i.e. LD_{50}, for pesticides. These models allow the assessment of the hazard of new pesticides without the use of animal testing, and to prioritise further experiments and guide future research.

Datasets used for the development of models are shown in the enclosed appendices.

CONTENTS

1. INTRODUCTION

1.1 PESTICIDES

Pesticides occupy an unique position among the many chemicals that man encounters daily, in that they are deliberately added to the environment to kill or injure some other form of life. The first synthetic pesticides became available during the 1940s, generating large benefits through increased food production. Concern about the adverse impacts of pesticides on the environment and human health started to be voiced in the early 1960s [1]. Since then, debate on the risks and benefits of pesticides has not ceased and a huge amount of research has been conducted into the impact of pesticides on the environment.

Ideally the toxic action of pesticides would be highly specific to undesirable target organisms and non-injurious to desirable, non-target organisms. In fact, however, most of the chemicals that are used as pesticides are not highly selective but are generally toxic to many non-target species, and other forms of life that co-habit the environment. Therefore, lacking highly selective pesticidal action, the application of pesticides must often be predicated on selecting quantities and manners of usage that will minimise the possibility of the exposure of non-target organisms to harmful quantities of these otherwise useful chemicals.

Each year an estimated 2.5 million tons of pesticides are applied to agricultural crops worldwide. The amount of pesticides coming in direct contact with, or consumed by, target pests is an extremely small percentage of the amount applied. In most studies the proportion of pesticides applied that reach the target species has been found to be less than 0.3%, so 99.7% went elsewhere in the

environment [2]. Since the use of pesticides in agriculture inevitably leads to exposure of non-target organisms, including humans, undesiderable side-effects may occur to some species, communities or to ecosystems as a whole.

Toxicological evaluation of the hazard relating to the handling and use of pesticides have, for many years, focused primarily on preventing injury to man. Common laboratory animals have served as the experimental models for man's biochemical, physiological, and pathological responses to these chemicals. Human pesticide poisoning and illnesses are the highest price paid when pesticides reach non-target areas. It is estimated that there about 1 million accidental human pesticide poisonings annually in the world, with about 20,000 reported deaths [3]. Taking into account both accidental and intentional (mainly suicide) exposures, the number of human pesticide poisonings has been estimated at about 3 million per year, with about 220,000 deaths [4].

In addition to concerns to human health there is increased awareness and concern for the ecological implications of the use of pesticides. This has started to direct the attention and research activities of toxicologists toward studies on wild species as well as to man, and the domestic and laboratory animals that are selected as test models to represent man. The study of the toxicology of pesticides, therefore, must take into account problems relating to both their harmful effects directly upon man and their effect on other species of animals in the environment from which man derives asthetic pleasure, as well as food, or which are essential to maintain a proper ecological balance. An increasing number of environmental effects of pesticides applications are being taken into account by regulatory bodies, leading to increased restrictions on the use of pesticides or to their ban. Although some of the environmentally most harmful pesticide uses have been eliminated, the options for pesticide use currently

available to farmers obviously differ with respect to the risk they pose to the environment.

Pesticides span a wide range of chemicals which differ in their chemical class and/or because they address a different type of pest. It is difficult to describe and cover all the possible types of pesticides, but some example of the best known classes are listed below:

- Organophosphate pesticides affect the nervous system by disrupting the enzyme that regulates acetylcholine, a neurotransmitter. Most organophosphates are insecticides. Some are very poisonous (they were used in World War I as nerve agents).
- Carbamate pesticides also affect the nervous system by disrupting an enzyme that regulates acetylcholine, a neurotransmitter. The enzyme effects are usually reversible.
- Organochlorine insecticides were commonly used in the past, but many have been removed from the market due to their health and environmental effects, as well as their persistence (e.g. DDT and chlordane).
- Pyrethroid pesticides were developed as a synthetic version of the naturally occurring pesticide pyrethrin. They have been modified to increase their stability in the environment. Some synthetic pyrethroids are toxic to the nervous system.
- Chlorophenoxy herbicides are sometimes mixed into commercial fertilizers to control the growth of broadleaf weeds. They are moderately irritating to skin, eyes, and respiratory and gastrointestinal linings.
- Nitrophenolic and nitrocresolic pesticides are highly toxic chemicals that have many uses in agriculture worldwide, as herbicides, acaricides,

nematocides, ovicides, and fungicides. They effects hepatic, renal and nervous systems.

- Arsenical pesticides, once absorbed, cause toxic injury to cells of the nervous system, blood vessels, liver, kidney, and other tissues.
- Biopesticides are derived from living systems, and generally are of a lower order of toxicity. *Bacillus thuringensis* is the most important live agent and specifically kills one, or a few related, species of insect larvae.

Many other specific agents, with widely varying toxicity are, or have been, produced in the past and a complete review of them is out of the scope of this thesis.

1.2 ENDPOINTS

1.2.1 Acute aquatic toxicity

The purpose of an acute toxicity test with fish species is to assist the assessment of possible risk to similar species in natural environments; as an aid in the determination of possible water quality criteria for regulatory purposes; and for comparative purposes by correlating with acute test results for other species. Data on a cold and warm freshwater species are generally required. The rainbow trout, *Oncorhynchus mykiss*, and the bluegill sunfish, *Lepomis macrochirus*, are the preferred species to meet this requirement. This is because they are sensitive indicator species and a large database which characterises their responses to environmental contaminants is available.

Testing of acute toxicity allows for the statistical estimation of concentrations that are expected to be lethal to a certain percentage of a group of organisms. The 50% concentration is the most commonly determined and is referred to as LC_{50}. Acute toxicity is measured after a relatively short exposure period, e.g. 96

hours, hence the 96h LC_{50}. LC_{50} measurements have been defined and standardised by procedures such as the OECD[1] [5] and/or EPA[2] [6] guidelines and then harmonised by OPPTS[3] [7].

1.2.2 Avian oral toxicity

Birds acquire toxic substances directly through their food and to a lesser extent through dermal exposure, preening, and grooming. However oral intake is considered to be the most significant route of exposure for free ranging species. As a consequence, testing for acute oral toxicity is one of the most important steps in determining the toxicological significance of any compound under investigation. Because birds are free to move in and out of areas which may have been recently treated with crop protection products, they run the risk of being exposed. The objective of oral toxicity testing is to determine the mean lethal dose (LD_{50}) that kills half (50%) of the animals tested. Test are usually conducted using either Northern Bobwhite Quail, *Colinus virginianus* or Mallard Duck, *Anas platyrhynchos*, however red winged black birds, house fiches, house sparrows and brown headed cowbirds can be used additionally, or optionally. Mortality is used as the primary toxicity endpoint; however, sub-lethal effects, such as changes in body weight reduced feed consumption, and changes in behaviour are carefully monitored and noted. Acute Oral Toxicity studies are based on the methods provided by the EPA [8]-[10], and OPPTS [11]. The study fulfills the data requirements for EU Council Directive

[1] Organisation for Economic Co-operation and Development.

[2] U.S. Environmental Protection Agency.

[3] Office of Prevention, Pesticides and Toxic Substances.

91/414/EEC, Anex II, 8.1.2, as amended by EU Commission Directive 96/12/EC.

1.3 QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS

Quantitative structure-activity relationships (QSARs) aim to find a relationship (model) between the chemical structures of compounds and a given activity. In computational chemistry, such a mathematical model is then responsible for representing and explaining the underlying mechanisms of the activity, for predicting activity values for other chemicals, and for designing new chemicals with a given activity value. More specifically in the field of toxicity, QSAR is used to derive predictive models for the impact of chemicals to human health, wildlife and environment.

However, a clear starting point for QSAR does not exist, but its roots developed during the 19th Century [12]-[16]. The earliest report on a relationship between molecular and biological properties seems to be published in a thesis by A.F.A. Cros, University of Strasbourg [17], in 1863. In 1868, A. Crum Brown and T. Fraser studied the biological effects of certain alkaloids, prior to and after methylation of a basic nitrogen atom. They observed pronounced differences between the basic and the permanently charged quaternary compounds, which led them to the conclusion that physiological activity should be a function of the chemical constitution [18]. Later, in the 1960s C. Hansch, T. Fujita [19], S. M. Free Jr. and J. W. Wilson [20] moved the science forward by a quantum leap, and started what is now considered to be classical QSAR. For a comprehensive review of modern QSAR refer to [21]-[26].

The basic assumption of QSAR is, of course, that a quantitative relationship between the molecular structure of compounds and their biological, chemical

and physical properties does exist. The development of QSARs arises from the interaction of a group of multidisciplinary experts from biology, chemistry and mathematics. In Figure 1 a schematic representation of the main building blocks used in development of a toxicity-based QSAR is outlined:

- a dataset that provides a measure of the activity for a group of compounds;
- the evaluation of the minimum energy conformation for each compound and the calculation of numerical descriptors related to each chemical structure;
- these two parallel data collections are then treated by some statistical technique that extracts relevant information to ascertain the underlying relationship between chemical structure and property. Finally the resultant QSAR should be validated and used only within the proper applicability domain of the model, to ensure that it is capable of providing sufficiently accurate predictions for new compounds.

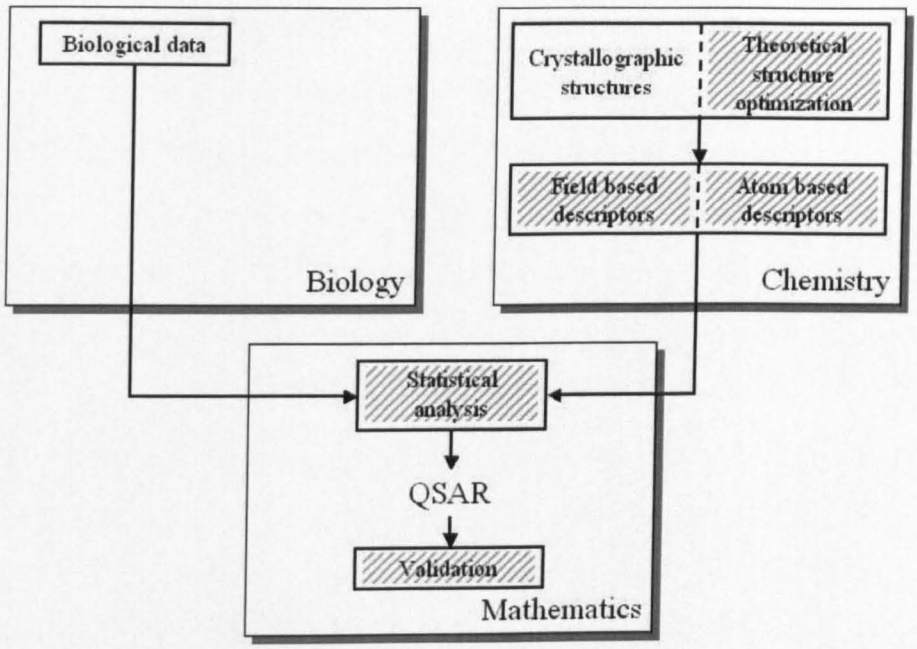


Figure 1. Building blocks used in the development of a toxicity-based QSAR. Hatched boxes are computer-assisted steps done by the QSAR modeller.

Guidelines and protocols are still an open issue in QSAR studies. Nevertheless, procedures defined by Hunger and Hansch in 1973 [27] are still topical and considered good practice in QSAR: select independent variables; justify the choice of the variables by statistical procedures; apply the principle of parsimony (Occam's razor); have a large number of objects, as compared to the number of variables; try to find a qualitative model of physicochemical or biochemical significance. Recently, in November 2004, these procedures where extended and integrated by the OECD for the validation of QSAR models for regulatory purposes [28]. An extract of the document is given below:

"[...] To facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information:

- a defined endpoint;
- an unambiguous algorithm;
- a defined domain of applicability;
- appropriate measures of goodness-of–fit, robustness and predictivity;
- a mechanistic interpretation, if possible [...]".

The intent of Principle 1 (defined endpoint) is to ensure clarity in the endpoint being predicted by a given model, since a given endpoint could be determined by different experimental protocols and under different experimental conditions. It is therefore important to identify the experimental system that is being modelled by the (Q)SAR. Further guidance is being developed regarding the interpretation of "defined endpoint". For example, a no-observed-effect level might be considered to be a defined endpoint in the sense that it is a defined information requirement of a given regulatory guideline, but cannot be regarded as a defined endpoint in the scientific sense of referring to a specific effect within a specific tissue/organ under specified conditions. The intent of Principle 2 (unambiguous algorithm) is to ensure transparency in the model algorithm that generates predictions of an endpoint from information on chemical structure and/or physicochemical properties. It is recognized that, in the case of commercially-developed models, this information is not always made publicly available. However, without this information, the performance of a model cannot be independently established, which is likely to represent a barrier for regulatory acceptance. The issue of reproducibility of the predictions is covered by this Principle, and will be explained further in the guidance material. The need to define an applicability domain (Principle 3) expresses the fact that (Q)SARs are reductionist models which are inevitably associated with limitations in terms of the types of chemical structures, physicochemical properties and mechanisms

of action for which the models can generate reliable predictions. Further work is recommended to define what types of information are needed to define (Q)SAR applicability domains, and to develop appropriate methods for obtaining this information. The revised Principle 4 (appropriate measures of goodness-of–fit, robustness and predictivity) includes the intent of the original Setubal Principles 5 and 6. The wording of the principle is intended to simplify the overall set of principles, but not to lose the distinction between the internal performance of a model (as represented by goodness-of-fit and robustness) and the predictivity of a model (as determined by external validation). It is recommended that detailed guidance be developed on the approaches that could be used to provide appropriate measures of internal performance and predictivity. Further work is recommended to determine what constitutes external validation of (Q)SAR models. It is recognised that it is not always possible, from a scientific viewpoint, to provide a mechanistic interpretation of a given (Q)SAR (Principle 5), or that there even be multiple mechanistic interpretations of a given model. The absence of a mechanistic interpretation for a model does not mean that a model is not potentially useful in the regulatory context. The intent of Principle 5 is not to reject models that have no apparent mechanistic basis, but to ensure that some consideration is given to the possibility of a mechanistic association between the descriptors used in a model and the endpoint being predicted, and to ensure that this association is documented.

1.3.1 Biological data

The first step in formulating QSARs is to build-up a data set that must reflect a well-defined toxic endpoint [21]. Data should be reliable since high quality toxicity data will have lower experimental error associated with them. All toxicity

measurements are subject to experimental error. The reality of toxicity testing is that even with a standardised protocol, it is not possible to obtain unbiased data, because different laboratory conditions, such as individual characteristics, age, and the health of test animals, will influence the results. Therefore, toxicity values are often reported as the mean for a series of replicates. A perfect statistical fit will never be achieved with a QSAR, and some degree of uncertainty in the model is expected and acceptable. Ideally the dataset should be designed to represent molecular diversity, but typically a few compounds tested are available.

When the database of toxicity values used to develop a QSAR becomes of sufficient molecular complexity, outliers will appear. Outliers are chemicals that do not fit the model, or are poorly predicted by it [29]. There are several potential reasons for a chemical being an outlier, generally such compounds can be recognised as acting by a different mode or mechanism of action from the other chemicals which are well modelled. They may also indicate poor data quality. Analysis of outliers in QSAR development assists in defining the applicability domain of the model.

1.3.2 Chemical structure representation

Chemicals are commonly thought of as a two-dimensional structures. However, their toxic effects are an expression of their three-dimensional structure. In living organisms the majority of biochemical processes follow the lowest-energy reaction pathway. Large biomacromolecules such as proteins, nucleic acids or polysaccharides, as well as small molecules such as peptides and hormones, normally exist in the most stable conformational state - the lowest energy conformation. In order to describe the 3D structural and electronic properties of

the molecule under consideration in a QSAR analysis correctly, one has to consider it in a stable (optimised) conformation.

Finding a global energetic minimum of a flexible molecule is not always an easy task. Even a small molecule composed of a few tens of the heavy (non-hydrogen) atoms containing a few rotatable bonds, has a high number of degrees of freedom. These include bond stretching, bond angle bending, and torsion angle rotations.

Throughout the recent decades of the development of QSARs and modelling methods, various techniques to obtain stable conformers of a compound have been found and applied in this research area.

- The most reliable molecular conformations can be found using experimental X-ray diffraction methods. These produce the "real" picture of the 3D arrangement of the heavy atoms of molecules in an analysed crystal. These methods are quite demanding due to the efforts necessary to grow the high quality crystals suitable for structural determination. The molecules in the crystals are packed and are subjected to intermolecular forces produced by surrounding molecules of the same compound, or the solvent, that may influence their conformation. However, the intermolecular forces are usually not too large and the conformation in the crystal is close to the global minimum. As X-ray crystallography is an experimental technique, the resolved structure can, in certain cases, be distorted due to the presence of multiple energetically accessible conformations of the same compound or impurities in the crystal. In this case, the molecular structure can be "fine tuned" by computational chemistry optimisation techniques. Crystal structures of biomacromolecules (e.g. proteins, nucleic acids) are stored within the Research Collaboratory for Structural Bioinformatics (RCSB)

database[4]. The Cambridge Crystal Database Centre[5] (CCDC) holds the structures of small organic molecules.

- Theoretical-computational approaches in conformational searching are based on the variation of the degrees of freedom in the molecule. The variations of torsion angles have the biggest impact on the energy of the molecule and determine the overall molecular shape. The values of the torsion angles can be varied either systematically or randomly. The most reliable results can be obtained with systematic variation of torsion angles, also called systematic conformational hyperspace sampling, in which several local minima as well as a single global minimum can be identified. The quality of the results depends directly on the capacity of the computational resources available. This is because the total number of conformations evaluated is determined by the number of torsion angles (rotatable bonds) searched in the molecule and the torsion angle variation step size. Due to this limitation, systematic conformational searching is usually used for small molecules with few rotatable bonds. With random searching methods the degrees of freedom of the molecule are varied randomly. A typical searching iteration consists of the random generation of the torsion angle values and geometry optimisation to the nearest local minimum. If the optimised geometry is unique, it is saved and used as a starting point for next iteration. The conformational search can finish if no new conformation can be found in a series of consecutive iterations.

[4] http://www.rcsb.org/

[5] http://www.ccdc.cam.ac.uk/

Random searching methods are suitable for the conformational analysis of large molecules, typically peptides and saccharides.

- The conformations obtained by X-ray crystallography or conformational searching are not absolute minima and, in most cases, need to be fully optimised. The main goal of the optimisation procedure is to minimise the internal strain of the molecule and find the nearest local, possibly global, minimum. The optimisation procedure is driven by a normal gradient, which reflects the energy changes with respect to structural changes. When the gradient value becomes low, the structure optimisation finishes at a minimum. The energy of the optimised molecule can be calculated by various methods of computational chemistry [30]. Molecular mechanics methods (e.g. MM2, MM+, AMBER) are based on the classical Newtonian laws. Atoms are approximated by balls and bonds by springs with a given force constant. The set of force constants for all atoms is called a "force field" and can be derived either from experiment (e.g. IR spectroscopy) or from a high level *ab initio* calculation. Molecular mechanics calculations are very fast and thus may be applied to study of molecular systems composed of thousands of atoms. In semi-empirical methods (e.g. CNDO, AM1, PM3, ZINDO) the computationally most expensive part of quantum chemical calculation (evaluation of two electron integrals) is either completely neglected or replaced by empirical parameters. This leads to increased performance in comparison to the full calculation. These methods provide good results for compounds that are similar to those used for parameterisation. Due to their good performance, they have been widely applied in QSAR. Finally *ab initio* quantum chemical methods are based on the first principles of physics and chemistry. These are the most expensive

computationally, but can provide structures and thermodynamic properties with an excellent agreement with experimental values.

1.3.3 Chemical descriptors

Having the structure in its minimum energy state, it is possible to calculate the descriptors that characterise the molecule mathematically. Since it is often difficult to know *a priori* the type of descriptor which might be relevant to the biological activity of interest, several (or many) parameters are calculated before starting any statistical analysis. Nowadays, commercial software packages have become available that possess the capability to calculate many hundreds of molecular descriptors from simple structural inputs with great ease (for a comprehensive compilation of QSAR parameters, with about 3000 references, see [31]). Because of this, a variety of properties have been used in structure-toxicity modelling. More often, the chemical descriptors used in structure-property correlations are based on the lipophilic, electronic and steric nature of substituents. The influence of molecular shape has always been difficult to describe. Despite this, a large variety of descriptors, from simple molecular weight to complex topological indices, have been employed to model steric properties. Of these, physicochemical descriptors of the hydrophobic, electronic and steric properties and quantum chemical values (including charges and orbital energies) have been consistently shown to be important in modelling toxicity. Although some of the steric descriptors, such as molecular volume, encode some 3D information, molecular conformation has not been considered. Descriptors can be divided according to type into two general classes.

Molecule-based descriptors

Molecule-based descriptors describe the magnitude of particular physical properties but do not consider the directional preferences that these properties may have. Molecular descriptors can vary greatly in their complexity. Some take the form of a binary indicator variable that encodes the presence of certain substructure or functional features. Other descriptors, such as HOMO (highest occupied molecular orbital) and LUMO (lowest unoccupied molecular orbital) energies, require semi-empirical or quantum mechanical calculations and are therefore more time-consuming to compute. Molecular descriptors are often categorised according to their dimensionality, which refers to the structural representation in which the descriptor values are derived. 1D-descriptors are generally constitutive (e.g., molecular weight). The 2D-descriptors include structural fragments, fingerprints or molecular connectivity indices. The molecular connectivity indices, which are based on graph theory concepts, can differentiate molecules according to their size, degree of branching, shape, and flexibility. As implied by the name, 3D-descriptors are generated from a three-dimensional representation of molecules. Some examples include molecular volume, solvent-accessible surface area, molecular interaction fields, or spatial pharmacophores. In addition to intrinsic dimensionality, molecular descriptors can be classified according to their physicochemical attributes. It is recognized that the dominant factors in receptor-drug binding are based on steric, electrostatic, and hydrophobic interactions. For many years medicinal chemists have attempted to model these principal forces of molecular recognition using empirical physicochemical parameters, which ultimately led to the introduction of fragment constants in early QSAR studies. These descriptors are constants

that account for the effect on a congeneric series of molecules of different substituents attached to the common core.

Field-based descriptors

Field-based descriptors describe the micro-environment surrounding the molecules and encode combinations of steric, hydrophobic and electrostatic properties, not only for molecular fragments, but for the whole molecule as well. The GRID [32] and CoMFA [33] programmes take advantage of molecular interaction fields by using different probe types (steric, electrostatic or lipophilic) in a 3D lattice environment. Other variants of the molecular field type, such as the molecular similarity-based CoMSIA approach [34], have also been reported in the literature. Most of the 3D-descriptors require a pre-aligned set of molecules. In cases where the exact molecular alignment is not obvious, one may consider the use of spatial invariant 3D descriptors (i.e., the descriptor values depend on conformation, but not spatial, orientation). CoMFA allows the modeller to ascertain a predictive relationship between molecular fields and biological activity. The model is expressed as the sum of contributions from every variable and the size of the coefficient for each variable that underlies the importance in describing activity. As each variable represents a variation in interaction energies at a defined point in 3D space, the regression coefficient can be mapped back onto the initial *x, y, z* coordinates of the variables, generating a 3D regression map. Therefore, the advantages of the CoMFA method are that it describes properties in terms of 3D fields, the results can be mapped into 3D space and one can localise points within the spatial distribution of properties which are related strongly to the activity. However, the major problem of CoMFA models stems from the alignment of compounds and, for

this reason, it is difficult to study highly heterogeneous datasets. A few innovative descriptors have been developed that do not depend on a mutual alignment of the molecules, e.g. grid-independent descriptors (GRIND) [35]. These are autocorrelation vectors of molecular surface properties that are independent of the relative orientation of the molecules in 3D space.

While it is generally accepted that toxicological assessments are subject to error, it should also be accepted that descriptor values used in QSARs are also subject to variability, and when possible the descriptors used in formulating the QSAR should allow for a mechanistic interpretation of the model.

1.3.4 Statistical analysis

Once biological data have been collected and chemical structures have been associated with a proper set of descriptors, mathematics is able to handle and extract the information hidden in the numbers.

The first step is to apply a statistical, or pattern recognition, method to correlate these descriptors with the observed biological activities. Partly due to the ease with which a great variety of theoretical descriptors may be generated, QSAR researchers are often confronted with high-dimensional data sets; the task in such a situation is to solve an ill-posed problem in which there are more variables (descriptors) than objects (compounds). The inflation of parameters posed the problem of variable selection. Topliss pointed out in 1972 that not only a large number of variables in the model, but also a large number of variables considered enormously increase the risk of chance correlation [36], [37]. The situation is even more complicated than it appears, because the underlying physicochemical attributes of the molecules that are correlated with their biological activities are often unknown, so that *a priori* feature selection is

not feasible in most cases. Thus, the selection of the best variables for a QSAR model can be very challenging. To reduce the risk of chance correlation and overfitting of data [38], the entire data set is usually pre-processed using a filter to remove descriptors with either small variance or no unique information. A feature selection routine then operates on the reduced data set and identifies which of the descriptors have the most prominent influence on the activity to build a model. There are two major advantages of feature selection. First, it can help to define a model that can be interpreted. Second, the reduced model is often more predictive, partly because of the better signal-to-noise ratio which is a consequence of pruning the non-informative inputs.

In the past, variable or feature selection was made by a human expert who relied on experience and scientific intuition. Other methods include the use of a correlation analysis of the data set, or by application of statistical methods such as forward selection or backward elimination. However, when the dimensionality of the data is high, and the interrelations between variables are convoluted, human judgment can be unreliable. Also, a simple forward or backward stepping algorithm fails to take into account information that involves the combined effect of several features, so that the optimal solution is not necessarily obtained [39], [40]. Recent developments in computer science have allowed the creation of intelligent algorithms capable of finding optimal, or near-optimal, solutions for such a combinatorial optimisation problem [41]-[48].

Having selected relevant features, the final stage of QSAR model building is executed by a feature mapping procedure. Among the numerous different techniques developed so far, regression-based analyses and classification techniques are prime examples in QSAR.

Regression-based analysis

Among these numerous techniques utilised to formulate a mathematical relationship the following are prime examples.

- **Linear methods:** Multiple linear regression (MLR) analysis was the traditional approach for QSAR applications in the past. MLR fits a regression model, $y = bX + c$, which models a response variable, y, as a linear combination of the X-variables. The major advantage of this method is its computational simplicity, offering the possibility to interpret the resulting equation easily. However, this method becomes inapplicable as soon as the number of input variables equals or exceeds the number of observed objects. As a rule of thumb, the ratio of objects to variables should be at least five for MLR analysis; otherwise there is a correspondingly large risk of chance correlation [37]. A common way to reduce the number of inputs to MLR, without explicit feature selection, is through feature extraction by means of principal component analysis (PCA) [49]. In this procedure, the complete set of input descriptors is transformed to its orthogonal principal components, relatively few of which may suffice to capture the essential variance of the original data. The new principal components are then used as the input to a regression analysis. Another very powerful multivariate statistical method for application to an underdetermined data set is partial least squares (PLS) [50], [51]. Briefly, PLS attempts to identify a few latent structures, or linear combinations of descriptors, that best correlate with the observations. Unlike MLR, there is no restriction in PLS on the ratio of data objects to variables, and PLS can deal with strongly collinear input data and tolerates some missing data values.

- **Non-linear methods:** traditionally, non-linear correlations in the data are explicitly dealt with by a predetermined functional transformation before entering a MLR. Unfortunately, the introduction of non-linear or cross-product terms in a regression equation often requires knowledge which is not available *a priori*. Moreover, it adds to the complexity of the problem and often leads to insignificant improvement in the resulting QSAR. To overcome this deficiency of linear regression, there is an increasing interest in techniques that are intrinsically non-linear. At the present time, artificial neural networks (ANN) are probably the most widely used non-linear methods in chemometric and QSAR applications. ANNs are computer-based simulations which mimic biological nervous systems. As in nature, they are composed of simple processing elements (neurons) operating in parallel (layers). The network function is determined largely by the connections between elements (weights), which are responsible for the network's intelligence. The inter- and intra-layer connections define the architecture of the ANN. ANNs can be trained to fit a particular function by adjusting the values of the weights between neurons. Commonly, ANNs are trained through a specific learning rule so that a particular input leads to a specific target output. Supervised learning is a recursive learning process where inputs fed in the ANN and are mapped in the output; the output is then compared with the target and network weights are adjusted accordingly until the network output matches the target. It has been demonstrated that multiple-layer neural network can approximate a continuous function to an arbitrary accuracy, given a sufficient number of neurons [52]. On the other hand unsupervised learning do not need target values, but they learn to recognise similarity among inputs.

Classification techniques

Examples of classification methods follow.

- **Discriminant analysis** is a technique for classifying a set of observations into predefined classes. The purpose is to determine the class of an observation based on a set of variables known as predictors or input variables. The model is built based on a set of observations for which the classes are known. Based on the training set, the technique constructs a set of linear functions of the predictors, known as discriminant functions, such that $L = b_1 X_1 + b_2 X_2 + \ldots + b_n X_n + c$, where the *b*'s are discriminant coefficients, the *X*'s are the input variables or predictors and *c* is a constant. These discriminant functions are used to predict the class of a new observation with unknown class. For a *k*-class problem *k* discriminant functions are constructed. Given a new observation, all the *k* discriminant functions are evaluated and the observation is assigned to class *i* if the *i*-th discriminant function has the highest value.
- **Decision trees** are built through a process known as binary recursive partitioning. This is an iterative process of splitting the data into partitions, and then splitting it up further on each of the branches. Initially all of the records in training set are together in one big box. The algorithm then tries breaking up the data, using every possible binary split on every field. The algorithm chooses the split that partitions the data into two parts such that it minimizes the sum of the squared deviations from the mean in the separate parts. This splitting or partitioning is then applied to each of the new branches. The process continues until each node reaches a user-specified minimum node size and becomes a terminal node. If the sum of squared

deviations from the mean in a node is zero, then that node is considered a terminal node even if it has not reached the minimum size.

- **Distance-based similarity analysis:** these methods require the definition of a distance function between objects and assume that it is possible to compute for each pair of objects in a domain their mutual distance (or similarity). In classical approaches, an object is represented by a point in a feature space. Then, the classification process consists in finding frontiers between classes. For each data point the classifier assigns it to the closest class in term of given distance or similarity.

1.3.5 Validation of QSARs

Hunger and Hansch [27] stated in1973 "One must rely heavily on statistics in formulating a quantitative model but, at each critical step in constructing the model, one must set aside statistics and ask questions. [...] without a qualitative perspective one is apt to generate statistical unicorns, beasts that exist on paper but not in reality. [...] it has recently become all too clear that one can correlate a set of dependent variables using random numbers as dependent variables. Such correlations meet the usual criteria of high significance". As such, model validation is a critical, but often neglected, component of QSAR development. In a recent review [53], Kövesdi and coworkers state that "[..] In many respects, a proper validation process is more important than a proper training. It is all too easy to get a very small error on the training set, due to the enormous fitting ability of the neural network, and then one may erroneously conclude the network would perform excellently". The first benchmark of a QSAR model is usually to determine the accuracy of the fit to the training data. However, because QSAR models are often used for activity prediction of

compounds not yet synthesised, the more important statistical measures are those giving an indication of their prediction accuracy. Common methods of choice to test QSAR predictivity are listed below.

- The most popular procedure for the estimation of the prediction accuracy is cross-validation, which includes techniques such as jack-knife, leave-one-out (LOO), leave-group-out (LGO) and bootstrap analyses [54]-[58]. The first group of methods is based on data splitting, where the original data set is randomly divided into two subsets. The first is a set of training compounds used for exploration and model building, and the second is the so-called validation set for prediction and model validation. The leave-one-out procedure systematically removes one data point at a time from the training set and, on the basis of this reduced data set, constructs a model that is subsequently used to predict the removed sample. This procedure is repeated for all data points, so that a complete set of predicted values can be obtained. It has been argued that the LOO procedure tends to overestimate model predictivity and that resulting QSAR models are over-optimistic [59]-[61]. However, the situation may be better in the case of large data sets, where cross-validation can be performed in larger groups [62], [63]. Technically, jack-knifing is used to estimate the bias of a statistic. A typical application of jack-knifing is to compute the statistical parameters of interest for each subset of data, and to compare the average of these subset statistics with the one that is obtained from the entire sample in order to estimate the bias of the latter. As an alternative to LOO, a LGO procedure can be applied which sets aside a percentage of the entire data set as a validation subset. In the literature, this procedure is also known as *k*-fold cross-validation, indicating that the entire data is divided into *k* groups of

approximately equal size. An added bonus of a LGO procedure is a vast reduction in computational resource relative to a standard LOO cross-validation. Bootstrapping represents another type of resampling method that is distinct from data splitting [64], [65]. It is a statistical simulation method which generates sample distributions from the original data set. The concept of bootstrapping is founded on the premise that the sample represents an estimate of the entire population, and that statistical inference can be drawn from a large number of pseudo-samples to estimate the bias, standard error, and confidence intervals of the parameter of interest. The bootstrap-samples are created from the original data set by sampling with replacement, where some objects may appear in multiple instances.

- Another popular means of statistical validation is the y-scrambling test [66], [67]. In this procedure, the output values, i.e. biological responses, of the compounds are shuffled randomly, and the scrambled data set is correlated by the QSAR method with the original *X* variables block. The entire procedure is repeated several, to many, times on differently scrambled data sets. If there remains a strong correlation between the descriptors selected and the randomised response variables, then the significance of the proposed QSAR model is regarded as being suspect.
- The real criterion for the validation of a QSAR model can only be good predictivity for an external test set which the model has never seen before. It is important that the compounds in the external test set must not be used in any manner during the model building process. Otherwise the introduction of bias from the test set compromises the validation process. Of course, the chemical space of the training and test sets must not be too different.

A variety of statistical parameters have been reported in the QSAR literature to reflect the quality of the model. These measures give indications as to how well the model fits existing data, i.e., they measure the explained variance of the target parameter y in the biological data. Some of the most common measures of regression are root mean squares error (rmse), standard error of estimates (s), and coefficient of determination (R^2). Generally, cross-validation performance significantly better (> 0.5) than that of y-scrambling tests, but not very different from that of the training set and external test predictions, is regarded as good trait of a robust and high-quality QSAR model. The evaluation of the goodness of a model in classification problems make use of error rate and misclassification matrix.

1.4 SUMMARY OF THE LITERATURE

Research community dedicated a lot of effort in the *in vivo* and *in vitro* studies of the toxicity and mechanism of action of the pesticides in aquatic organisms. Among most recent works we can cite [68]-[71].

Unfortunately, in spite of the large number of QSAR published in literature, few work has been specifically addressed to the study acute aquatic toxicity of pesticides. A pioneer work has been done by Mager, in 1982, who shown that the neurotoxicity of organophosphorus pesticides depended on lipophilic and steric substituent properties using the response surface optimisation of the MASCA model [72], [73]. Recently, two general QSAR models for predicting the acute toxicity of pesticides to *Oncorhynchus mykiss* [74], and *Lepomis macrochirus* [75] have been published. Rather than being real predictive models of acute toxicity of new pesticides, authors studied, using multivariate techniques, in particular a three-layered back-propagation neural network, the

influence of weight of fish, time of exposure, and experimental conditions, such as temperature, pH, and hardness of the water, on toxicity. In particular they found that toxicity increase with the time of exposure and/or temperature; on the other hand, the weight of fish seems to have only a limited influence on the toxicity of the pesticides.

The research on the effects of pesticides to bobwhite quail has been primarily focused on *in vivo* tests. Some publications demonstrated that some pesticides can mimic vertebrate steroids, and interact with steroid receptors and reproductive function in quails [76]-[78] or induce changes in hepatic microsomal enzyme systems [79]. The cholinesterase activity of pesticides has been studied in few works [80], [81]. But a statistical analysis of the relationship between the chemical structure of pesticides and their avian toxicity devoted to build-up a predictive model is still missing.

2. SCOPE OF THE WORK

The specific objectives of this thesis can be summarised as follows:

- to develop a model for the prediction of acute aquatic toxicity;
- to describe mathematically the mechanisms of avian oral toxicity;
- to evaluate and verify the capabilities of QSAR as a tool for the prioritisation of further experimental tests.

3. MATERIALS AND METHODS

This section gathers together and describes tools and techniques developed during the PhD study and that were exploited for the development of QSAR models.

Paragraph 3.1 lists the sources used for collecting the ecotoxicological data and describes the protocol used for ensuring their quality.

The OpenMolGRID (Open Computing Grid for Molecular Science and Engineering) system (paragraph 3.2) is a unified and extensible information-rich environment for solving molecular design/engineering tasks relevant to chemistry, pharmacy and life sciences. The OpenMolGRID system comprises a set of application-oriented tools that are built on core Grid services and functions provided by the UNICORE infrastructure. This system is the main outcome of a EU founded joint project (contract: IST-2001-37238).

The following paragraph (3.3) describes how the OpenMolGRID system was used for 3D optimisation of the structures of chemicals, and for the calculation of the chemical descriptors.

In paragraph 3.4 it is presented a proprietary algorithm, based on genetic algorithm which was used in following sections for the selection of relevant variables.

This section ends with a paragraph (3.5) dedicated to the description of the rationale used for the development of QSAR models.

3.1 ECOTOXICITY VALUES

The data sets used were extracted from the US EPA-Office of Pesticides Programs (EPA-OPP), SEEM (produced by the International Center for

Pesticides and Health Risk Prevention), and BBA (Federal Biological Research Center for Agriculture and Forestry) ecotoxicology databases for aquatic toxicity data.

The database of EPA-OPP was developed to make readily accessible an up to date summary of the EPA reviewed data corresponding to the ecotoxicological effects of all pesticides active substances presently registered or previously manufactured in the US. Data have been produced according to the US-EPA guidelines, and checked to ensure compliance with the guidelines and data quality.

ICPS (international Centre for Pesticide and Health Risk Prevention) produced a database of ecotoxicological endpoints within the SEEM project (Statistical Evaluation of available Ecotoxicology data on plant protection products and their Metabolites, sponsored by the EC, DG Health and Consumer Protection, contract n° B1-3330/2001216). The lists of endpoints from the ECCO peer review process and from the national review process were selected for the project, as the two sources of available validated data. Literature was searched as a complementary source.

Data in the BBA database are a collection of endpoints from studies conducted for regulatory purposes by the Federal Biological Research Centre for Agricultural and Forestry (BBA).

In order to ensure good high quality for QSAR development, criteria for consistency and reliability of the data has been applied, such as compliance to standardized procedures such as OECD [5] and EPA guidelines [6], or to Good Laboratory Practice (GLP), or availability of ancillary data such as purity, year of the study, uncertainty of the experimental result, and other statistical parameters.

QSARs refer to well defined chemical compounds, which are described using 2D/3D descriptors or fragments. Therefore data referring to mixture of chemicals were discarded. However, mixtures of stereoisomers were kept, because they are super imposable using common 2D descriptors.

Then, studies with an active substance < 85% purity, and data given as higher or lower that values were discarded.

A final issue is the selection of a single toxicological value when more than one is available. To address this point the following scheme has been applied (Figure 2). The first step verifies the presence of multiple values. In this case we evaluated the value spread. If the highest value was more than four time the minimum, the compound was discarded, as conflicting results appeared. If more than one value existed, within a factor of four, the minimum was used, choosing among the studies defined by the US-EPA as core studies. Studies were assigned to this category if they fulfil the basic requirements of current guidelines and are acceptable for use in a risk assessment.

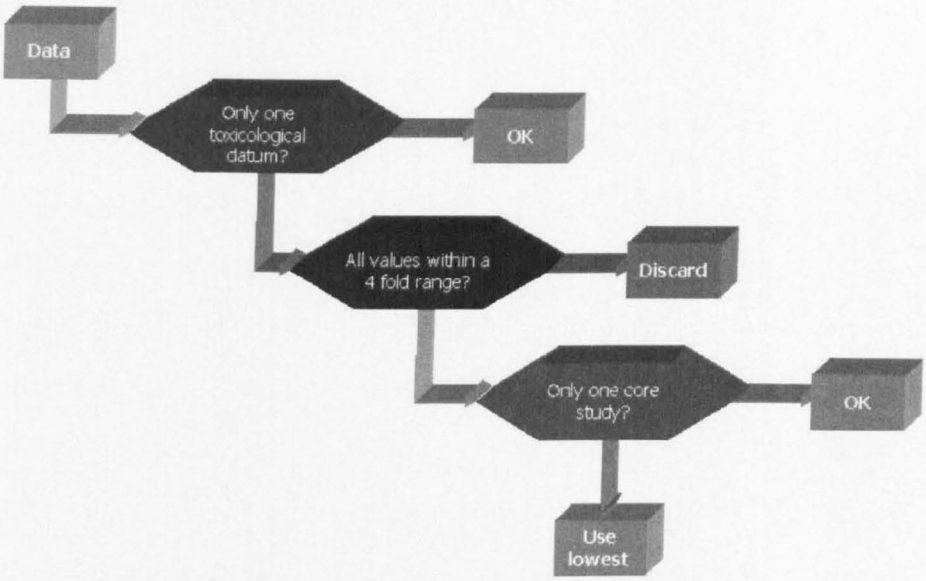


Figure 2. Scheme adopted to select toxicological data from databases.

The adopted procedure guarantee high quality for several reasons:

- the quality arises from the compliance to the EPA protocols;
- hierarchical screening eliminated less reliable data on the basis of additional general criteria such as reference of purity of the active substance and invalid studies;
- internal and external comparison of databases allowed for the selection of robust data in case of multiple values.

For a detailed description of this protocol the reader is referred to [82].

The collection of the ecotoxicological data was done by the Laboratory of Environmental Chemistry and Toxicology at the Mario Negri Institute within the DEMETRA project (www.demetra-tox.net).

3.2 THE OPENMOLGRID SYSTEM

The OpenMolGRID system is a joint effort of the OpenMolGRID consortium (www.openmolgrid.org).

3.2.1 Introduction

The use of *quantitative structure-property/activity relationship* (QSPR/QSAR) methods responds to the need of pharmaceutical and chemical companies to screen *in silico* millions of potential new drugs or chemicals. They also assist in the need of research institutes and regulatory bodies to have fast, accurate and reliable models to understand and predict the consequence of chemicals to human health, wildlife, and the environment. The basic concepts, together with new approaches in QSPR/QSAR, have been reviewed several times [83]-[86]. In a recent paper [21], Schultz and Cronin identified the essential and desirable characteristics of quantitative structure-activity relationships for ecotoxicity. These can generally be applied to every QSPR/QSAR method. According to these characteristics, the development of QSPR/QSARs should be based on:

- a reliable dataset, which differs both in terms of potency and chemical structure;
- a set of descriptors of superior quality, which are reproducible, and of a number and type so as to be constant with the property being modelled, and when possible, allow for a mechanistic explanation;
- a rigorous and appropriate statistical process; and
- a strict validation procedure of the model.

For the purpose of general models, following these four rules, researchers easily may end up with an extremely large dataset to be analysed, and in spite of the potential capabilities of modern computers, the computational effort for

such studies will be massive. The well-known axiom *time is money* applies also to QSPR/QSARs and the calculation procedure has to be reduced in terms of computational time. The new Grid technology seems to be adequate to address such a challenge because of its ability to exploit distributed computational resources.

The specific components of OpenMolGRID are discussed below. The general structure and technology of the project is described in Section 3.2.2. Section 3.2.3 outlines the data warehousing component, which is designed to integrate information from disparate locations. The choice and the adaptation of existing QSPR and QSAR analysis software for the Grid environment is the subject of Section 3.2.4. Section 3.2.5 contains conclusions and an outlook on further applications of the OpenMolGRID system.

3.2.2 OpenMolGRID architecture

The OpenMolGRID project has several requirements in terms of the underlying Grid infrastructure: (1) Existing computational software packages need to be integrated, with particular emphasis on support for complex, multi-step workflows. (2) Computationally intensive tasks need to be executed in a distributed fashion to reduce turn-around times. (3) Access to heterogeneous data sources is needed, where the strict security requirements of the pharmaceutical industry need to be taken into account. (4) And above all, the user interface of the system has to be as user-friendly as possible, with most of the Grid-related complexity hidden from the user, while still providing all of the flexibility and power of the underlying Grid system for advanced users.

The UNified Interface to COmputing Resources (UNICORE) [87] Grid infrastructure was chosen as the foundation of the OpenMolGRID system.

Briefly UNICORE can be characterised as a vertically integrated Grid system, with an emphasis on seamless and secure access to Grid resources. It offers a powerful and easy-to-use graphical user interface, single sign-on, and strong security through X.509 public key cryptography. The UNICORE plug-in interface allows for the straightforward integration of new applications. OpenMolGRID uses the open interfaces provided by UNICORE to integrate novel applications such as databases and software packages. Figure 3 shows the general structure of the OpenMolGRID system.

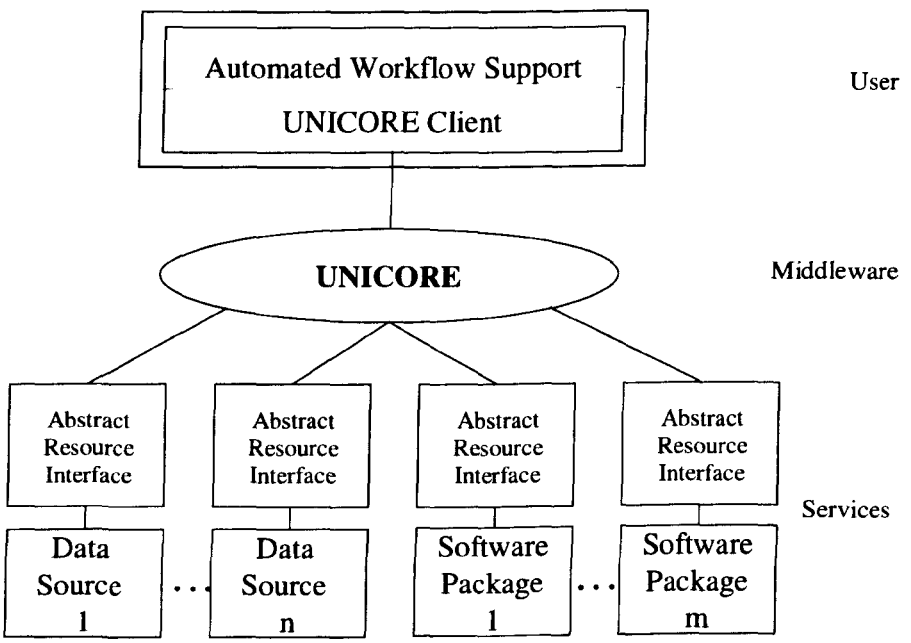


Figure 3. General OpenMolGRID architecture.

OpenMolGRID extends UNICORE, adding significant new functionality in the areas of workflow support and resource management. On the client side, a new type of UNICORE plugin, called MetaPlugin, supports the user in dealing with

complex workflows. It enables users to build UNICORE jobs from abstract workflow descriptions, where details such as file transfers, dependencies and resource allocation are taken care of automatically. Computationally intensive tasks can be run on multiple sites, if the input data can be split into smaller pieces and distributed. The resources needed for the job are identified and allocated automatically by the MetaPlugin.

On the server side, an abstraction layer, called the Abstract Resource Interface, is used to access software resources in a generic fashion. Data sources are integrated in the same general way using an Abstract Resource Interface. An important task of the Abstract Resource Interface is to allow the use of standardised input/output formats, thus creating an abstraction layer around the underlying software package.

The extension of the UNICORE infrastructure, that is required to provide the functionality offered by the OpenMolGRID system, is performed in an application domain in an independent, flexible and extensible fashion by using an XML-based metadata layer.

All Grafical User Interface (GUI) plugins for new software packages developed within OpenMolGRID are also useable as standalone components, thus creating added value for UNICORE, even without taking advantage of the full OpenMolGRID system.

3.2.3 Data warehousing

The molecular engineering process of the OpenMolGRID system is supported by data mining tools and systems. Data mining techniques, such as *multi-linear regression* (MLR), *principle component analysis* (PCA), *partial least squares* (PLS), and *artificial neural networks* (ANN), are used to build predictive

QSPR/QSAR models [83]. Data warehousing is often employed as a pre-requisite to data mining [88]. A data warehouse integrates, cleanses, normalises, and consolidates data from different sources and maps them onto "ready-to-use" data structures (e.g. by de-normalizing relational database tables). The main function of the *OpenMolGRID data warehouse* (MOLDW) is to harvest chemical-compound data from public resources and integrate and pre-process them for the data mining and molecular engineering process within OpenMolGRID. The diagram in Figure 4 illustrates the basic logical structure of the MOLDW and its relationship to other system components.

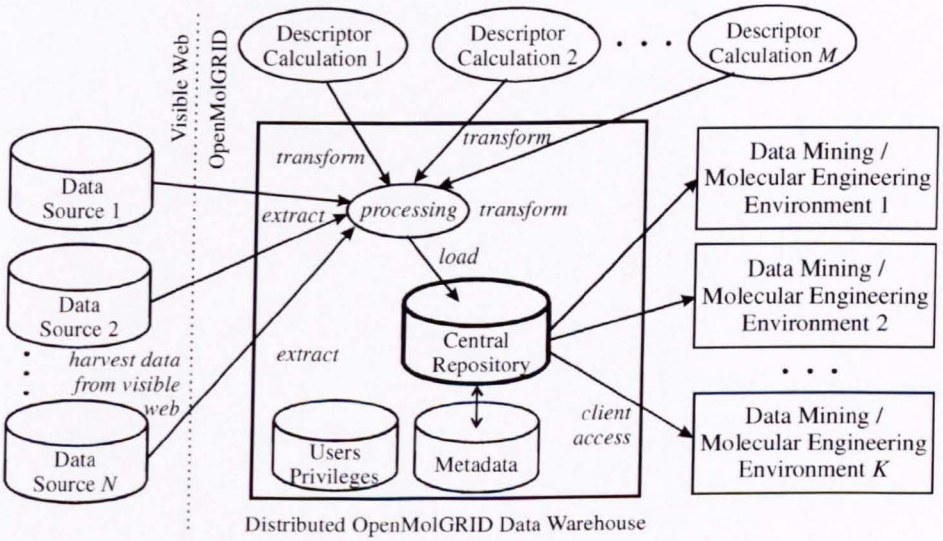


Figure 4. MOLDW and other OpenMolGRID system components.

What is interesting from a Grid perspective is the requirement (1) to harvest data from public repositories into a protected Grid computing environment (the OpenMolGRID), and (2) to incorporate physically distributed data transformations, so-called *descriptor calculations*, into the logically integrated data warehouse. Descriptor calculations are fundamental to *in silico* molecular

modelling. Some descriptors are calculated from the three-dimensional structure information, e.g. molecular volume, quantum chemical descriptors. Specialised software is required to perform these calculations and typically they are expensive to compute, especially if there are a large number of chemicals and several representations of the same chemical. Clearly, when the data warehouse is updated, MOLDW will only update an entry and re-compute descriptors if the entry in the underlying database has been modified, avoiding needless computation. MOLDW effectively "caches" computations (i.e. stores the results of computations) and thus facilitates more efficient data mining downstream, as it removes the burden from data miners to carry out the required integration and transformations.

Currently, various aspects of MOLDW and its interoperation, both with the visible web and the OpenMolGRID system, have been designed and implemented. These include the logical and physical data model of the central storage (realised on a PostgresQL relational database platform), the database access tool (DBAT), and certain aspects of the ELT (extract, load, transform) [88] processes. Some more advanced warehouse access and query tools (e.g. fingerprinting, substructure search) are subject to being developed in the near future.

3.2.4 Modules for QSPR/QSAR modelling

The QSPR/QSAR models are designed by finding relationships between property/activity and molecular structures. This process involves various tasks that are carried out at different stages of a complicated workflow. A typical workflow starts with the extraction of a training set with the experimental property/activity values from the data source (e.g. data warehouse, database,

file system). Normally, the prerequisite for the model development is the calculation of molecular descriptors [31], [89], which are used to represent molecular structures in the model. The descriptor calculation itself can be a multi-step process and depend on the generation of 3-dimensional coordinates and performing quantum-chemical calculations. Currently, thousands of different molecular descriptors are available and various data mining techniques (MLR, PCA, PLS, ANN etc) can be used to select the significant descriptors that have causal relationships with the modelled property or activity.

Each of the tasks described above can be performed with different software packages that are often incompatible with each other. However, the most optimal design of the predictive models requires the combined application of multiple software packages. This problem is addressed within the OpenMolGRID infrastructure by the development of UNICORE compliant applications that adapt existing software modules for the design of predictive QSPR/QSAR models. These OpenMolGRID applications can be then combined to carry out complex workflows. As described above, each application consists of two parts – the plugin to the UNICORE Client and the Abstract Resource Interface. This architecture allows different software packages to be used when performing one specific task in the workflow.

A set of programs are integrated into OpenMolGRID to demonstrate its capabilities:

- **2D to 3D conversion:** The MOLGEO software [90] has been adapted for the conversion of 2D structures to 3D representations. This is a common data pre-processing task in QSPR/QSAR modelling, since the 2D representation is very convenient to the end-user for sketching molecular structures and as most chemical databases have only 2D representations

available. However, all quantum chemical, and most molecular descriptor, calculation programs require the 3D representation of molecular structures as an input.

- **Semi-empirical quantum chemical calculations:** The MOPAC (version 7) [91] software has been adapted for semi-empirical quantum chemical calculations. MOPAC is a general-purpose semi-empirical quantum mechanics package for the study of chemical properties and reactions in gas, solution or solid-state. The output from MOPAC calculations is used to calculate quantum-chemical descriptors (e.g. dipole moment, heat of formation, energy partitioning, reactivity indexes, etc.) for QSAR/QSPR model development.
- **Descriptor calculation:** The MDC module from the CODESSA PRO software has been adapted for molecular descriptor calculation. Currently, the system incorporates a wide range (about 1000) of molecular descriptors, describing constitutional, topological, structural and electronic features of structures. The descriptor calculation module is applicable both to 2D and 3D structures, although 3D structures provide a more information rich description of the molecules.
- **Model development:** The MDA module from the CODESSA PRO [92] software has also been adapted for QSAR/QSPR model development. Multiple statistical methods are available for the development of predictive models, including Multilinear Regression Models (MLR) and Partial Least Squares (PLS). Several algorithms are available for the selection of descriptors to search effectively for the best (most informative) multi-parameter correlations in the large space of natural descriptors. The predictive capability of the model is judged by statistical parameters

calculated for the model, various cross-validation techniques, internal and external validation sets. Visualisation tools are available for plotting actual vs. predicted activities/properties and residuals

In addition, new molecular engineering tools have been developed for the computer-aided construction of molecular structures with predefined chemical properties or biological activities. These tools will make it possible to explore large chemical space in a cost effective way to find potential candidates for new drugs, chemicals, or materials. The generation of new molecular structures is based on a library of fragment structures. Using that library, various structure generation algorithms can construct a huge number of candidate structures. The candidate structures are then validated using the previously developed predictive models and a small subset of molecules that match the target properties or activities is selected for further investigation.

3.2.5 Conclusions

OpenMolGRID is one of the first realisations of Grid technology in drug design. The system is designed to create QSPR/QSAR models and use them to predict biological activities or absorption, distribution, metabolism, and excretion (ADME) related properties. OpenMolGRID is based on the adaptation and integration of existing, widely accepted, relevant computing tools and data sources, using the UNICORE infrastructure, to make a solid foundation for the next step molecular engineering tools. Using the implemented data warehouse technology, the system is suitable to collect data from geographically distributed, heterogeneous sources.

The system is capable of solving molecular engineering problems on a large-scale. In particular the system facilitates the discovery of novel compounds with favourable properties by analysing millions of structures in a reasonable time.

3.3 CALCULATION OF DESCRIPTORS

Chemical descriptors were obtained using the OpenMolGRID system. The OpenMolGRID project is developing tools for secure and seamless access to distributed information and computational methods relevant for molecular engineering. A full description of the system can be found earlier in this thesis (section 3.2) or in [93] and [94]. For this study the workflow shown in Figure 5 was used. It consists of the following five steps:

- 2D structure sketches (connectivity formule) were entered into the OpenMolGRID system;
- structures were converted into three dimensions using MOLGEO 1.0 [Algorithm settings: distance geometry; Tolerance: 3; Time limit: 10; Add Hydrogens: ON];
- 3D structures were optimised by MOPAC 7.05 using the gradient criterion [Keywords: AM1 T=3600 NOINTER MMOK GNORM=0.1 EF];
- single point calculations of thermodynamic properties at optimised geometries were performed using MOPAC 7.05 [Keywords: AM1 VECTORS BONDS PI POLAR PRECISE ENPART MMOK 1SCF];
- optimised 3D structures of the compounds, with additional thermodynamic output files for each structure, were used as the input for the CODESSA PRO software. This calculates a large pool of constitutional, geometrical, topological, electrostatic, surface area, quantum-chemical, molecular orbital, and thermodynamic descriptors.

Values of the logarithm of octanol/water partition coefficient (LogP) were calculated by KowWin1[6] [95], and added to the dataset because of its well known relevance in predicting aquatic acute toxicity.

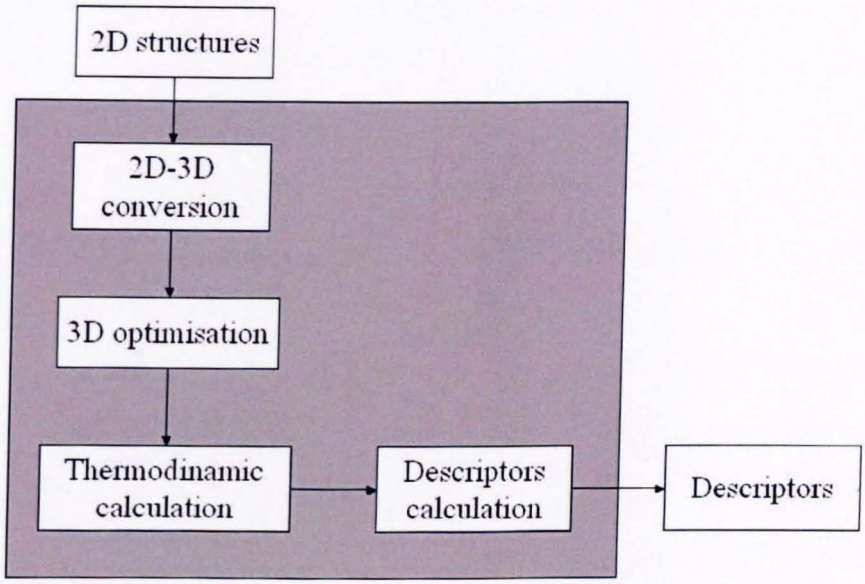


Figure 5. Descriptor calculation workflow: OpenMolGRID system integration is indicated within the grey box.

3.4 GENETIC ALGORITHM

The study of QSARs relates a property to the corresponding structures of the training set. Consequently, provided with a method to describe structures, a good predictive model can be constructed. Hence, it is beneficial to generate large numbers of descriptors containing topological, geometric, electronic and quantum-chemical features that maximise the amount of information in the input

[6] http://www.syrres.com/esc/kowwin.htm.

space. However, as the information content spreads over a very large number of potential molecular descriptors, it remains difficult to exploit. It is well known that increasing the number of variables will often cause a reduction in the generalisation ability of the model (the curse of dimensionality) and some QSARs based descriptors do not add information, but only increase noise. Therefore, it is necessary to select a subset of descriptors that retain most of the intrinsic information content. A number of mathematical and statistical methods [42], [44], [45], [96]-[101], even if they prove to be efficient in some applications, rapidly exhibit limitations in large datasets [102], [103]. Therefore the use of these techniques in QSAR is not suitable.

This current work proposes a general methodology to search a solution space based on genetic algorithms (GAs) for hyperspace exploration. These approach is presented as a preliminary test bed for future application in QSAR studies. GAs have already been considered to solve general optimisation problems [102], [104]-[106].

3.4.1 Materials and Methods

A GA is a stochastic global search method that mimics natural biological evolution [104], [106], [107], [108]. GAs operate on a population of potential solutions, applying the principle of survival of the fittest to produce approximations to a solution. At each generation, a new set of approximations is created by the process of selecting individuals according to their level of fitness in the problem domain and breeding them together using operators from natural genetics. This process leads to the evolution of populations of individuals from a seed population, just as in natural adaptation. Individuals, or current approximations, are encoded as strings, *chromosomes*, composed over some

alphabet(s), so that the *genotypes* (chromosome values) are mapped uniquely onto the decision variable (*phenotypic*). In the context of variable selection, the representation is the binary alphabet {0, 1}, where 0 defines the absence of the descriptor, and 1 defines its presence.

A flow-chart of the procedure is shown in Figure 6. The population at time *t* is represented by the time-dependent variable *P*, with initial population of random estimates being *P(0)*. Using this outline, the remainder of this section describes the major elements of the procedure.

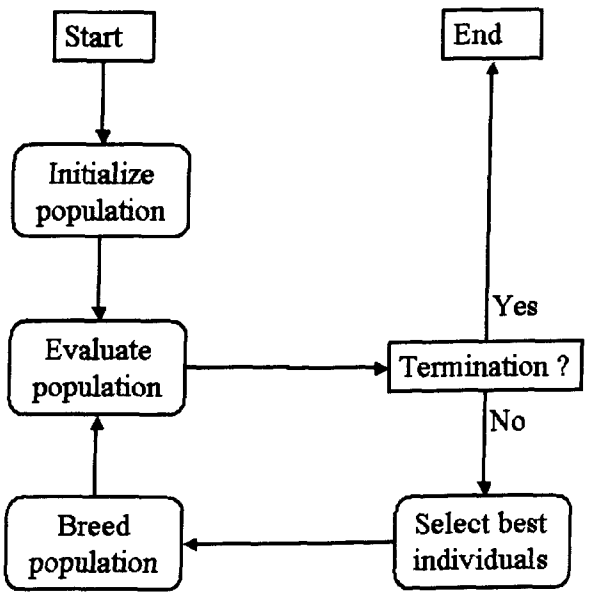


Figure 6. Flow-chart of GA.

Population representation and initialisation

The first step of GAs is to create an initial population. This is achieved by generating the required number of individuals using a random number generator that uniformly distributes numbers in the desired range. The structure

of each chromosome is composed of *m* variables, representing its dimension, each coded by a bit, 0 or 1; where 0 means the variable is not active, while 1 indicates its activity.

Evaluation of the population and objective function

The objective function is used to provide a measure of how individuals perform in the problem domain. In a minimisation problem, the fittest individuals will have the lowest numerical value of the associated objective function. The index proposed to evaluate the performances of the individuals is the performance of the QSAR model is being optimised.

For a regression problem the objective value proposed is the root mean squared error (rmse) as follows:

$$r = y - \hat{y},$$

$$rmse = \sqrt{\left(\frac{\sum r^2}{n}\right)}$$

$$ObjF = rmse$$

where y is the experimental value, is the predicted value, and n is the number of data points.

For a classification problem the objective value to be minimised is the error rate (ER).

Selection and reproduction procedure

The chromosomes are evaluated using the objective function previously defined (1), and only the best individuals are retained for the reproduction procedure. Selection is the process of determining the number of trials that a particular individual is chosen for reproduction and, thus, the number of offspring that an individual will produce. The algorithm supports two different mechanisms to select individuals: a stochastic universal sampling function, *SUS* [109], and the

"roulette wheel" selection, *RWS* [104]. SUS is a single-phase sampling algorithm with minimum spread and zero bias. RWS is a stochastic sampling with replacement. A description of the procedure implemented is in the user's guide of the Genetic Algorithm Toolbox [110]-[112].

Crossover is the basic operator for reproducing new chromosomes in the GA. It produces new individuals that have some parts of both parents' genetic material. In this work, multi-point crossover [113] (Figure 7), m crossover positions, k_i $\{1, 2, \ldots, l - 1\}$, where k_i are the crossover points and l is the length of the chromosome, are chosen at random with no duplicates and sorted. Then, the bits between successive crossover points are exchanged between the two parents to produce two new offspring. The disruptive nature of multi-point crossover appears to encourage the exploration of the search space, rather than favouring the convergance to highly fit individuals early in the search, thus making the search more robust.

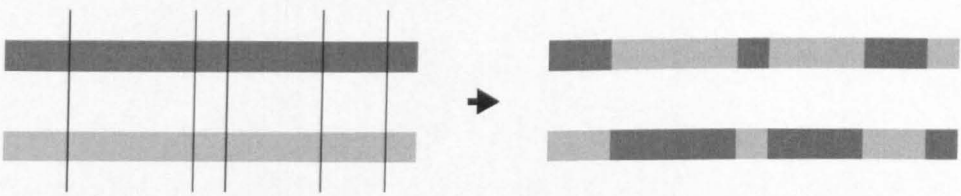


Figure 7. Multi-point Crossover (m=5) in a GA selection procedure.

In natural evolution, mutation is a random process where one allele of a gene is replaced by another to produce a new genetic structure. Finally, mutation (Figure 8) is randomly applied with low probability and modifies elements in the chromosomes. When chromosomes are represented by a binary string, as in this case, the mutation operator randomly switches some bit. The mutation serves to create random diversity in the population and it should prevent the algorithm converge towards a non-optimal solution.

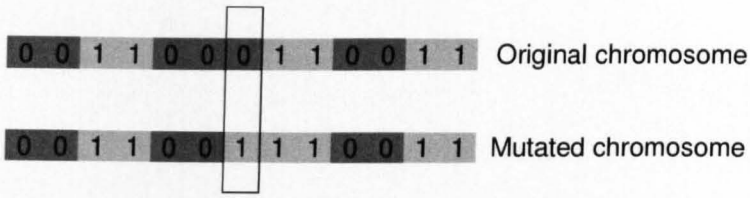


Figure 8. Binary mutation.

Termination

Because the GA is a stochastic search method, it is difficult to specify convergance criteria formally. As the fitness of a population may remain static for a number of generations before a superior individual is found, the application of conventional termination criteria becomes problematic. A common practice is to terminate the GA after a specified number of generations and then check the nature of the best members of the population. If no acceptable results are found, or the chromosome population is not comparable, the GA may be restarted or a fresh search initiated.

Software and computational details

The proprietary software was developed in MATLAB® (The MathWorks, Natick, MA) using elements from the Genetic Algorithm Toolbox [110], [111], [112] of Chipperfield et al. (Department of Automatic Control and System Engineering, University of Sheffield, Western Bank, Sheffield, UK). For this study the algorithm was implemented on a Intel® Pentium® III Mobile CPU 1200MHz processor.

3.4.2 Test of the method

The procedure was tested for regression problems using counterpropagation neural network for deriving the fitness score (GA/CPNN). CPNN is described in details later in this thesis (paragraph 4.1.2).

Artificial data

The GA was first tested on five well defined artificial data sets, described below. Outliers from a QSAR are chemicals that do not fit the model or are poorly predicted by it. Outliers are always present in experimental data sets and are useful in QSAR development as they assist in establishing the chemical domain of the model. Therefore, in order to simulate real experimental data sets, some outliers were added to the artificial dataset.

Table 1. CPNN parameters for the artificial data set.

CPNN parameter	Value
Dimension of the layer	10
Type of neighbourhood corrections	Triangular
Maximal correction factor	0.50
Minimal correction factor	0.01
Epochs	1000

For this analysis the parameters listed in Table 1 and Table 2 were used. These parameters allowed steady and repeatable runs. Figure 9a-e shows the five target functions.

Table 2. GA parameters for the data sets.

GA parameter	Artificial data	Academic data set I	Academic data II
Chromosome number	5	10	12
Chromosome size	10	34	50

Number of generation	100	250	300
Mechanism of selection	SUS	SUS	SUS
Rate of individuals to be selected	1	1	1
Probability of crossover	0.7	0.7	0.7
Crossover point number	2	2	2
Probability of mutation	0.02	0.02	0.02

Data set I

For this data set one vector of 100 random numbers uniformly distributed between 0 and 1, x_1. was generated. Nine vectors of normally distributed random numbers (zero mean, unit variance) were also generated to simulate Gaussian noise. The response was modelled as $_1 = f(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10})$. Variables x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9 and x_{10} represent only Gaussian noise. The true model is:

$$y_1 = x_1$$

Five more randomly generated objects were added to this data set in order to simulate outliers. As an example, an extract of the data set I is shown below:

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	y_1
1	0.950	0.226	0.934	0.916	0.829	0.013	0.276	0.907	0.217	0.970	0.950
2	0.231	0.580	0.264	0.602	0.166	0.310	0.368	0.759	0.652	0.715	0.231
...	...	...	...	...	...	...	...	...	...	...	...
100	0.988	0.393	0.740	0.784	0.170	0.069	0.196	0.930	0.233	0.065	0.988
101	0.583	0.592	0.432	0.986	0.540	0.853	0.787	0.310	0.008	0.375	0.253

102	0.423	0.120	0.634	0.473	0.623	0.180	0.619	0.269	0.397	0.374	0.585
103	0.516	0.038	0.803	0.903	0.686	0.032	0.016	0.536	0.650	0.484	0.524
104	0.334	0.459	0.084	0.451	0.677	0.734	0.891	0.163	0.085	0.969	0.163
105	0.433	0.870	0.945	0.805	0.877	0.537	0.762	0.211	0.769	0.342	0.486

This data set allowed the evaluation of the real ability of the GA to explore 10-dimensional variable hyperspace in the presence of noisy information. The system was able to recognise the correlation between the actual variable x_1 and pointed out as irrelevant the remaining nine variables. The function is represented in Figure 9a.

Data set II

A set of independent variables was simulated by generating nine vectors of 100 random numbers distributed uniformly between 0 and 1. One further vector was added as Gaussian noise, x_{10}. The target function is a linear combination of nine variables:

$$y_2 = x_1 + x_2 + x_3 + x_4 + x_5 - x_6 - x_7 - x_8 - x_9$$

Five outliers were added to this data set. The function is displayed in Figure 9b. In this case, the presence of numerous relevant variables makes the exploration of the hyperspace a difficult task. For this particular problem a backward elimination could probably achieve the aim faster because only one of the variables should be discarded, but the problem of stopping the procedure still remain. The GA correctly interpreted 90% of the variables, in fact all the variables but x_4 were correctly identified.

Data set III

For this data set, the same procedure as before was used to simulate five variables, with uniformly distributed random numbers between 0 and 1, and five variables representing Gaussian noise. In this case the complexity of the response y is slightly increased (Figure 9c):

$$y_3 = x_1 + 2x_2 + 4x_3 - x_4 - 3x_5$$

Again, five outliers were added. The target function was still linear, but the importance and influence of each variable in the response was different. The system correctly distinguished nine variables and misclassified only one variable (x_1).

Data set IV

For this data set, 2 vectors of 100 random numbers were again generated and were used as input variables. Eight vectors of 100 random numbers that were used as Gaussian noise. The true model is:

$$y_4 = x_1^2 - \log_{10}(x_1) + 3x_2$$

In this case the efficiency of the objective function, i.e. CPNN, was analysed. The dependence of the response y_4 to the variables x_1 and x_2 is very complex and highly non-linear. The system correctly selected all the actual variables.

Data set V

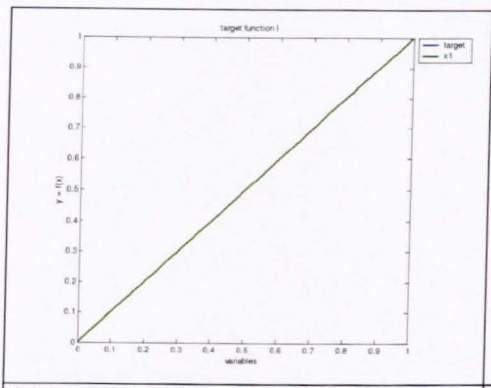
This data set was generated as previously stated. Figure 9e displays the true model:

$$y_5 = x_1 + x_2^2 + \log_{10}(x_3) + \frac{1}{(x_4 + 1)} - 2x_5$$

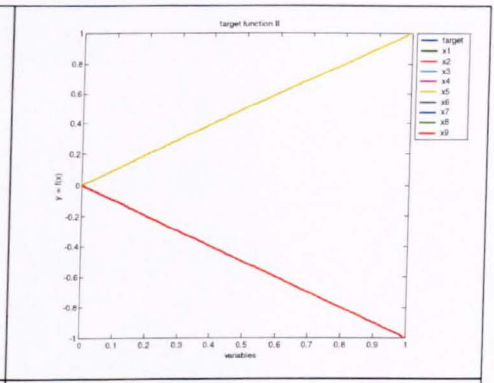
Five more objects were generated randomly and added to the data set as outliers. In the final test five variables (x_1, x_2, x_3, x_4, x_5) are involved in the definition of the response y_5 and five variables represent only Gaussian noise.

The target function presents both linear and non-linear correlation. This makes this analysis a strong and reliable test for both hyperspace exploration and chromosome selection. In this case it correctly interpreted 80% of the variables, i.e. x_2 and x_4 were not selected as relevant variables.

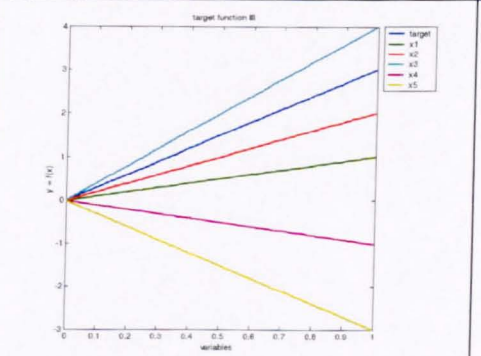
Results of the analysis for all target functions are summarised in Table 6.



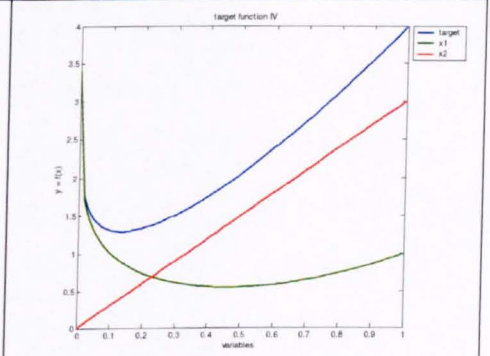
(a) target function I



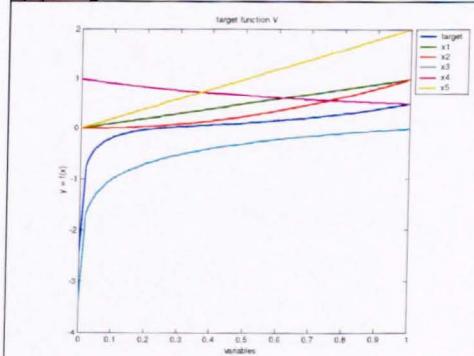
(b) target function II



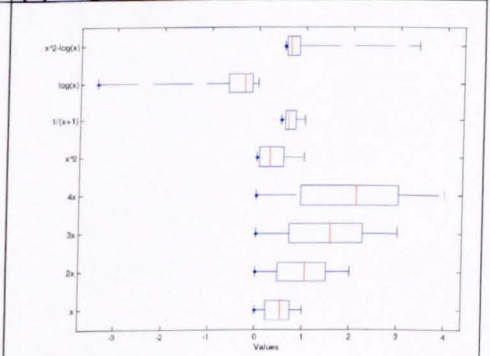
(c) target function III



(d) target function IV



(e) target function V



(f)

Figure 9. (a) transformed variables in target function I. (b) transformed variables in target function II. (c) transformed variables in target function III. (d) transformed variables in target function IV. (e) transformed variables in target function V. (f) statistical information of the transformed variables: the box is the interquartile range (the difference between the 75^{th} and 25^{th} percentile of the data); the line in the middle of the box is the sample median; the lines extending the box show the extent of the rest of the sample.

Academic data

The method was also evaluated on real academic data sets, which allowed the extension of the analysis and the conclusions to actual data sets for QSAR studies.

Academic data set I

This data set contains 92 compounds and the corresponding chronic dose rate to the mouse that would give half the animals tumours (TD_{50}). The original data set was collected by Gold and colleagues [114] and contains more than 1200 chemicals. The compounds to be evaluated were limited to those containing an aromatic ring and a nitrogen linked to the aromatic ring. Chemicals with no carcinogenic effect to the mouse were discarded. The data set was supplemented by 34 molecular descriptors as described in previous work [115]. The output (TD_{50}) was transformed as follow:

$$y_6 = \log\left(MW \cdot \frac{1000}{TD50}\right)$$

where *MW* is the molecular weight, and then normalised between 0 and 1 using a range scaling procedure, in order to have a more continuous output space and refer to molar units rather than by weight [132].

From the 34 descriptors calculated, 15 were selected using the parameters listed in Table 1 and Table 2: these were HOMO, LUMO, heat of formation, dipole moment, Randic Index, Wiener Index, Kier & Hall connectivity index order 0, Kier & Hall connectivity index order 4, log D at pH 2, log D at pH 7.4, first principal moment of inertia, third principal moment of inertia, Kappa simple index 1, Kappa alpha index 2, and electrotopological sum (descriptors are described in the original work [115]).

Using the parameters listed in Table 1 and 46 molecules random selected as training set, two models were developed exploiting firstly all 34 descriptors and then the 15 selected descriptors. The models were then tested on the remaining 46 molecules. The determination coefficient (R^2) for the test set exhibited a considerable improvement using the selected descriptors (see Figure 11 and Table 5).

Academic data set II

Debnath et al. [116] collected the mutagenic activities of a set of aromatic and heteroaromatic amines in *S. typhimurium* TA98 with a S9 microsomal preparation. Basak and Mills [117] supplemented the data set by molecular descriptors including topostructural, topochemical, geometrical and quantum chemical indices. This data set contains 95 chemical compounds with their mutagenic activities and 50 variables.

The parameters listed in Table 1 and Table 2 were used to select a subset of relevant variables through GA/CPNN and, as a result, 26 descriptors were selected: ^{3D}W, $^{3D}W_H$, kp0, kp1, kp2, kp3, LUMO, ^{4}v, $^{5}v_{Ch}$, $^{10}v_{Ch}$, $^{5}b_{CP}$, SIC_4, I_{ORB}, ASZ_5, SHCsats, Sumdel*I*, SHsOH, NumHBd, Gmin, SddsN, NHBint9, SssNH, I^{W}_{D}, ^{4}, $^{4}{}_{PC}$, DSN_1 (descriptors are described in the original work [117]).

The model developed exploiting the subset obtained showed a significant increase in the predictive ability of 47 random selected objects not included in training set over the model including all the variables (Figure 11 and Table 5).

3.4.3 Discussion

Synthetic data

Figure 9 shows the transformed variables X_i in the target functions I (a), II (b), III (c), IV (d) and V (e). Transformed variables are the variables derived from the original considering the true model to be intrinsically linear. For instance, in the target function III the transformed variables for x_1, x_2, x_3, x_4, and x_5 are:

$$X_1 = x_1 \quad X_2 = 2x_2 \quad X_3 = 4x_3 \quad X_4 = -x_4 \quad X_5 = -3x_5$$

Figure 9f displays some common statistical measures for the transformed variables, as range, interquartile range (the difference between the 75th and 25th percentile of the data) and median.

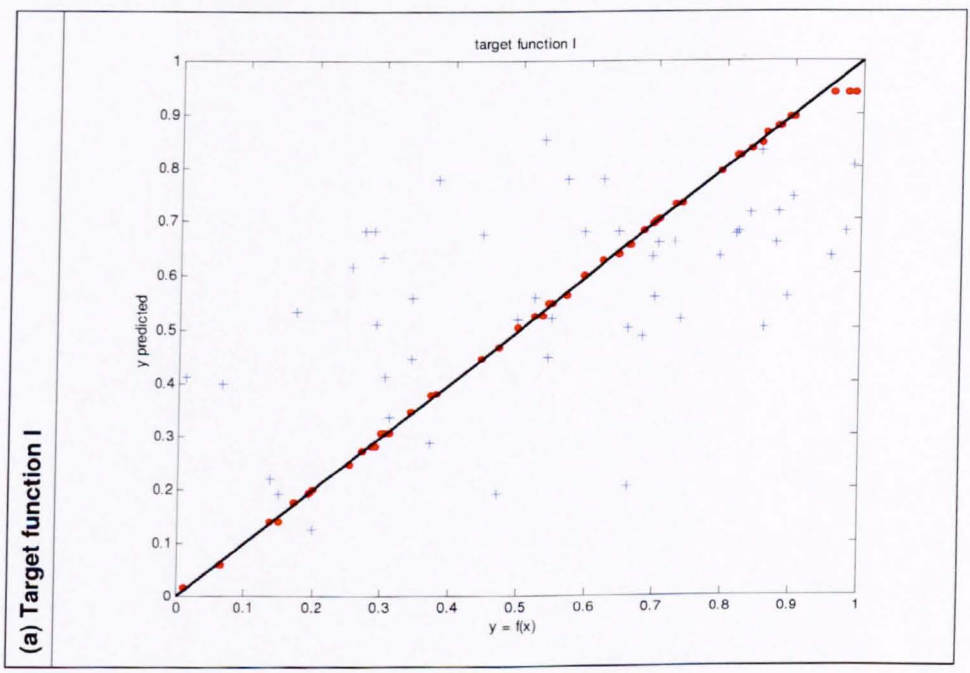
Analysing these plots (Figure 9) it is possible to comment on the errors in the selection of the variables. In the target function V (Figure 9e and Figure 9f) X_2 and X_4 have the lowest variability in respect of the other transformed variables. In this case the role of these input variables, i.e. x_2 and x_4, in the target function is not as important as the other variables and the procedure is not able to detect their correlation with the response y_5. Moreover, if Figure 9e is studied carefully it is observed that the transformed variables X_2 and X_4 are the inverse of the other in the observed interval, making their contribution similar to a constant (std = 0.1821, mean = 1.0363).

In the case of target function II (Figure 9b) all the actual variables have the same importance in the response y_2; the difficulty in selecting x_4 is probably related to its particular distribution. In fact, if two different variables are "similar"

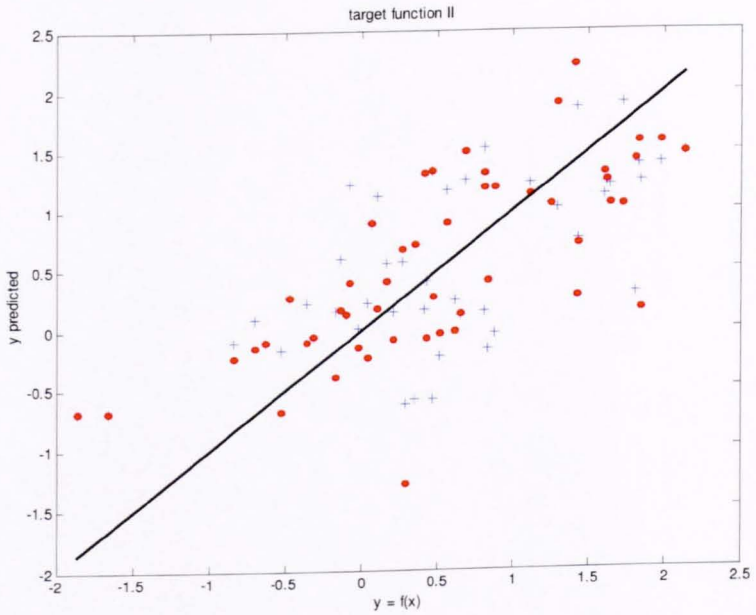
or correlated, i.e. $x_{a,i} \cong x_{b,i}$ for $\forall\ i$ where x_a and x_b are vectors/variables, the procedure may recognise only one of them. For instance, if x_a and x_b are "similar" vectors, the response $y = x_a + x_b$ can be easily mistaken for $y = 2x_a$ or $y = 2x_b$.

For the target function III similar considerations apply. X_1 and X_4 have the lowest variability and the contribution of x_1 to the response y_3 is probably merged with the linear contribution of the other variables.

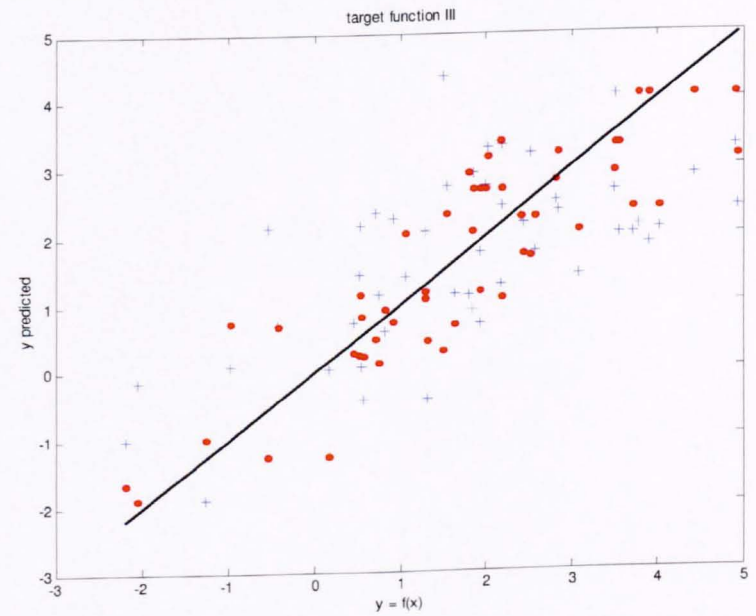
Models were developed using a CPNN trained with and without variable selection and in all the cases the selection of the variables showed a significant improvement in modelling performances (Figure 10).



(b) Target function II



(c) Target function III



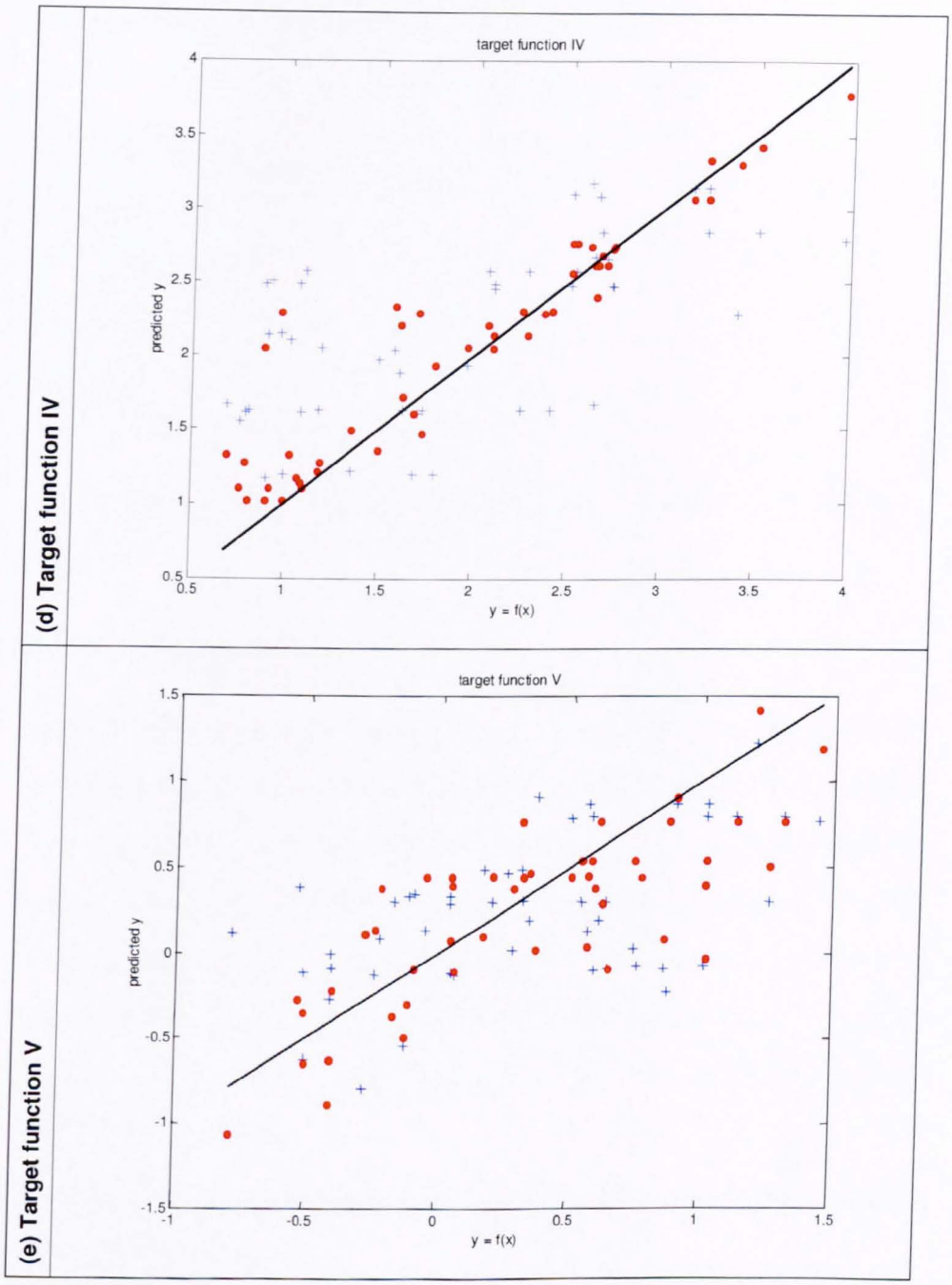


Figure 10. Prediction of the test set by a model developed using all the variables (+) and only the variables selected (.) for the target function I (a), the target function II (b), the target function III (c), the target function IV (d) and the target function V (e).

Table 3. Determination coefficient (R^2) of the test sets.

	I	II	III	IV	V
R^2 using all the variables	0.295	0.455	0.419	0.379	0.277
R^2 using the variables selected	0.999	0.525	0.765	0.868	0.414

A full exploration of the variables hyperspace would involve the generation of *t* models,

$$t = \sum_k \left(\frac{p!}{k!(p-k)!} \right)$$

where *p* is the total number of variables and *k* is the maximum dimension of the model, i.e. maximum number of variables involved in the model. In this case, *p* = 10 and *k* = 10, it would require the generation of 1023 models. But the exploitation of the GA by this procedure allowed the generation of only 500 models (Table 2). The performances are then compared to the results of Multiple Linear Regression Analysis (MLRA). The performances of a linear approach can be easily foreseen by simply *eye-balling* the correlation matrix for each model data (Table 4).

Table 4. Correlation coefficients of independent variables to dependent variables in the artificial datasets.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
y_1	0.976	0.080	0.004	0.111	-0.121	-0.179	-0.013	-0.078	-0.179	-0.167
y_2	0.433	0.337	0.404	0.451	0.168	-0.464	-0.436	-0.224	-0.397	-0.082

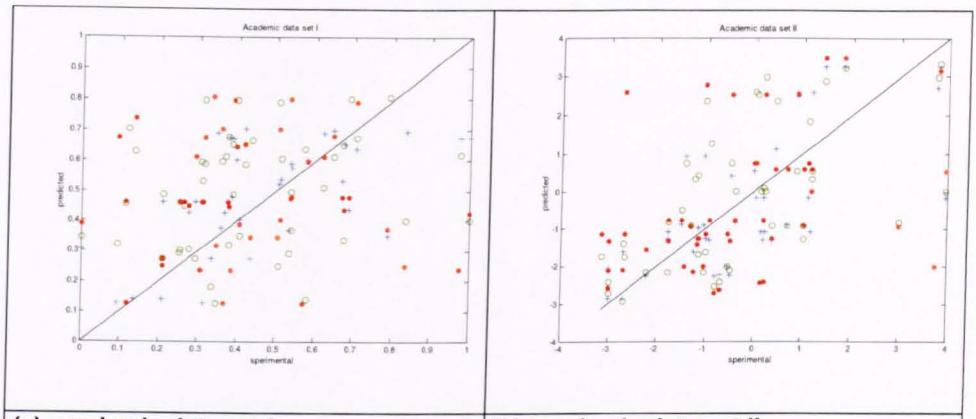
y_3	0.260	0.365	0.614	-0.038	-0.524	-0.106	-0.107	-0.107	-0.160	0.126
y_4	-0.045	0.859	-0.110	0.027	-0.029	-0.136	-0.067	0.087	-0.070	0.020
y_5	-0.161	-0.052	-0.106	-0.354	-0.060	0.114	0.133	0.053	0.046	-0.084

For this analysis an arbitrary choice of cut-off value of the correlation coefficient to determine whether a variable is, or is not, relevant was used. Knowing the true model and observing Table 4 it was observed that no relevant variable have an absolute correlation coefficient below 0.2. Setting the cut-off value to 0.2 means all the information about the data sets is used and the best possible selection is made. However, even in this most fortunate case, this analysis misclassified: x_5 in data set II; x_4 in data set III; x_1 in data set IV; x_1, x_2, x_3, x_5 in data set V (Table 6).

Academic data

In this case a comparison with respect to the true model is not possible. Therefore the method was evaluated simply on the predictive ability of the model developed exploiting the selected variables. This was compared with similar models developed using all variables, and variable selection from traditional methods. It is important to note the superiority of the method proposed as regards to traditional methods.

For the first data set, principal component analysis (PCA) was used to select a smaller set of descriptors [115] and then a similar CPNN was trained using the same parameters and the same molecules for the training set. The predictive ability of the remaining test set shows an improvement in the performances (R^2 = 0.038) but was not comparable with the selection obtained by GA/CPNN (R^2 = 0.297) (Figure 11a and Table 5).



(a) academic data set I | (b) academic data set II

Figure 11. Prediction of the test set using all variables (.), variables selected using GA/CPNN (+), and variables selected by other methods (o) for the academic data set I (a) and the academic data set II (b).

In the second case the selection was obtained using regression methodologies. Within the numerous variable selections performed by Basak and Mills [117] that which gave the best results in their models was chosen. Again, similar CPNN were trained using the same parameters and the same object and then tested on the remaining molecules of the data set. Once again the selection obtained by GA/CPNN showed the best results (Figure 11 and Table 5).

Table 5. Determination coefficient (R^2) of the test sets.

	R^2 using all variables	R^2 using variables selected by GA/CPNN	R^2 using variables selected by other method
I	0.001	0.297	0.038
II	0.404	0.637	0.603

3.4.4 Conclusions

Table 6 summarises the results of the analysis for the artificial data sets. The Non Error Rate percentage (NER%) was computed considering the number of variables, out of the whole pool, correctly interpreted by GA/CPNN as relevant or not. It is important to underlie that the identification of an irrelevant variable as being irrelevant is also a correct identification.

Table 6. Overview of the results from the artificial data set.

	Description	Actual variables	Noisy variables	Variables selected by GA	GA NER%	Variables selected by MLRA	MLRA NER%
I	Linear correlation	x_1	x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}	x_1	100	x_1	100
II	Linear correlation	x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9	x_{10}	x_1, x_2, x_3, x_5, x_6, x_7, x_8, x_9	90	x_1, x_2, x_3, x_4, x_6, x_7, x_8, x_9	90
III	Linear correlation	x_1, x_2, x_3, x_4, x_5	x_6, x_7, x_8, x_9, x_{10}	x_2, x_3, x_4, x_5	90	x_1, x_2, x_3, x_5	90
IV	Non-linear correlation	x_1, x_2	x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}	x_1, x_2	100	x_2	90
V	Non-linear correlation	x_1, x_2, x_3, x_4, x_5	x_6, x_7, x_8, x_9, x_{10}	x_1, x_3, x_5	80	x_4	60

A new algorithm suitable to select relevant variables from a problem domain has been explored. This algorithm, derived from GA concepts for hyperspace

exploration was combined with a CPNN to derive a specific score index, evaluating the quality of the selection. The fitness score of each chromosome was derived by the determination coefficient of the test set. This strategy can slow down the speed of convergence of the algorithm but assures the selection of descriptors subsets that lead to general and suitable models, by preventing over-fitting.

The selection power of the proposed method was tested on artificial data sets. 100 objects described by 10 variables were generated and relationships formed by five different target functions to the response *y*, five objects were added and used as outliers. The method allowed the derivation of relevant subsets of descriptors in all cases (Table 6). Two models were developed for each target function (Figure 10); the first one was trained without descriptor selection and the second one was developed by exploiting the descriptors selected by the procedure. The examination of the results confirmed immediately that the GA selection procedure allowed a notable improvement in the predictive ability of the models. Its would be noted that GA/CPNN was able to recognise linear and non-linear relationships between variables and the response in data sets holding both poor and abundant information.

The method was also tested on real literature data sets for toxicity. The performance of the subset of variables obtained with GA was compared with the variable selection obtained in the previous work on the same data sets. Rough models were developed under the same conditions, but the variables used. In both cases the selection resulting from the method gave the best results.

Finally, it is known that it is time consuming to select variables using the genetic algorithm approach than that using other methods. However, when the

descriptor pool is large, as in the case of QSAR studies, the advantages of using a genetic algorithm will be distinctive and significant.

3.5 RATIONALE FOR MODELLING

In a recent publication [118], Lemke et al. gave a neat mathematical description of QSARs for ecotoxicity. Ecotoxicological tests observe the evolution in the time of systems that interact strongly with each other and in which the behaviour is not expressible as a linear function of its descriptors. Or, using a mathematical expression, they observe a time-variant non linear dynamic system, that is outlined as in Figure 12. Particularly in the case of aquatic acute tests, the ecological system is the pool where the concentration of the examined pollutant, *p*, varies and the biological system is the population of observed fishes.

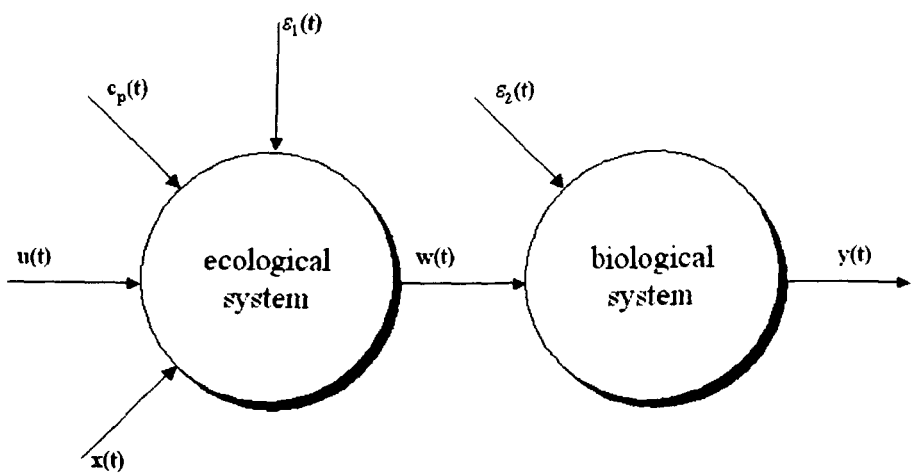


Figure 12. Dynamic model of an ecotoxicological system.

In Figure 12, $x(t)$ is the state vector of the ecological system at time t, $u(t)$ is the vector of external variables at time t, $c_p(t)$ is the concentration of the pollutant p at time t, $\varepsilon_1(t)$ and $\varepsilon_2(t)$ are the external disturbances to the system at time t, and $y(t)$ is the output vector describing the health of the population at time t. During the experimental tests, however, the external variables $u(t)$ and the state variables $x(t)$ of the ecological system are not usually observed or not observable and therefore considered constant. Consequently, the system reduces to a static system, that can be represented by the following non-linear static model (Figure 13):

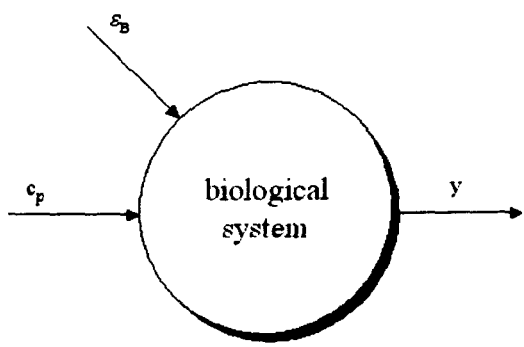


Figure 13. Reduced model of the static system measured in toxicological tests.

where $\varepsilon_B = h_1(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4)$ are the external disturbances to the system, ε_3 is the additional noise introduced to the static system by missing information of the

external and state variables that now transform to noise, and ε_4 is the noise added by the testing procedure itself. This system can be described by a function that takes the form of $y = f_1(c_p, \varepsilon_B)$.

On the other hand, QSAR studies aim at finding a description of the dependence of a chemical's property, in this case LC_{50}, or any other biological activity or effect from the chemical's molecular structure s_p (Figure 14):

$LC_{50} = f_2(s_p, \varepsilon_M)$, with ε_M

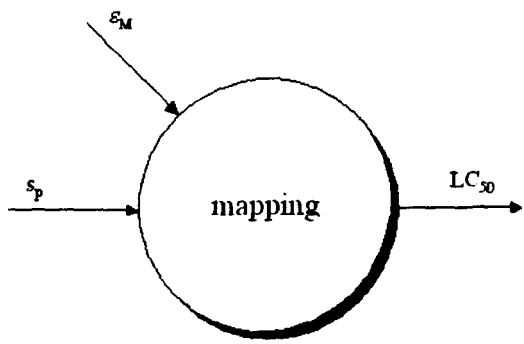


Figure 14. The QSAR problem. Note that the input variable c_p (LC_{50}) of the initial ecotoxicological system (Figure 12 and Figure 13) has shifted to being the objective of modelling.

A next problem is how to express the structure s_p of the chemical p. Commonly, it is a complex chemical object, but for building a mathematical model that describes the dependence of the toxicity from the chemical structure

a formal transformation into a set of numerical properties, i.e. descriptors, is required. This transformation is based on chemical and/or biological domain knowledge implemented in some software (Figure 15):

$d_p = f_3(s_p, \varepsilon_T)$.

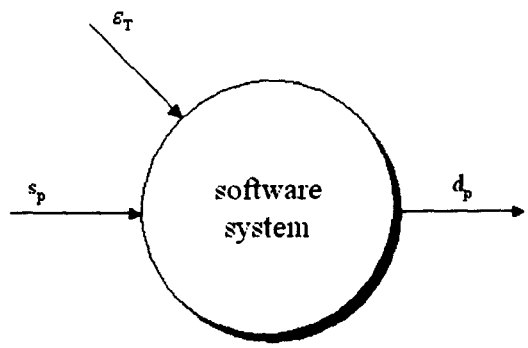


Figure 15. Model of the chemical description.

The process of descriptor calculation also add noise, ε_T. Not only software bugs or manual failures may introduce noise, more important for introduction of uncertainty should be interpretation clearance of domain knowledge for properly formalising an appropriate set of molecular descriptors, different starting condition assumptions (conformation) for descriptors calculation, or several different optimisation options.

The final, simplified non-linear static model used in QSAR modelling to describe acute toxicity is $LC_{50} = f_2(f_3(s_p, \varepsilon_T), \varepsilon_M) = f_4(s_p, \varepsilon_T, \varepsilon_M)$, (Figure 16):

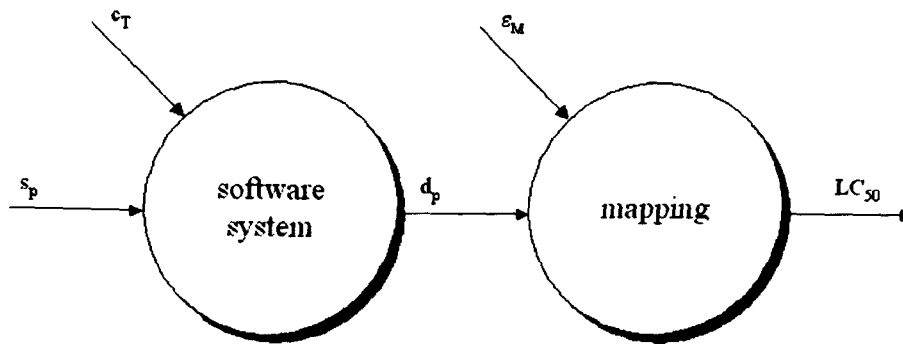


Figure 16. Complete QSAR model.

where, LC_{50} is the experienced lethal concentration causing a lethal effect in 50% of test animals for a certain species and chemical compound; s_p is the structure of the tested chemical compound in the chemical domain; ε_T is the noise of the chemical structure to molecular descriptor transformation process; ε_M is the noise transformed from the ecotoxicological system; d_p is the vector of numerical molecular descriptor of the test compound.

The external disturbance ε_T, which adds noise to descriptor input space used for modelling, can be reduced by fixing bugs and manual failures and by finding the most consistent chemical structure to descriptor transformation. However it is not clear *a priori* which transformation or optimisation will add, and which will reduce, noise. The disturbance ε_M, which finally results from the experimental

tests, in contrast, adds noise to the output LC50 and is a given fact that cannot be changed afterwards.

It is therefore difficult to extract relevant information about the dynamic and mechanisms of the ecotoxicological system using the approximations implicit in a QSAR approach. Nevertheless, some general deduction can be done if the inevitable noise generated during the procedure is minimised. To do so, the following steps were taken:

- quality checks are applied to screen the experimental data collected;
- the molecular descriptors generation is automated and defined;
- a hierarchical approach, where models of increasing computational complexity are used in a graduated manner, is adopted;
- the knowledge gained by a QSAR study is validated thoroughly by appropriate statistical methods.

4. ACUTE AQUATIC TOXICITY

This chapter of the thesis summarises the work on the development of a QSAR model for the prediction of the acute aquatic toxicity of pesticides.

4.1 MATERIALS AND METHODS

4.1.1 Dataset

Data for the rainbow trout (*Oncorhynchus mykiss*) acute toxicity (LC_{50}) following an 96h exposure were screened as described previously (paragraph 3.1). The protocol adopted, provided reliable ecotoxicity values for 282 compounds. As common and good practice in ecotoxicological QSARs, data were then transformed and modelled as $Log_{10}(1/LC_{50})$ [mmol/L].

The chemical structures of these compounds were then optimised and supplemented by descriptors as described in paragraph 3.3. Initially, eight compounds were excluded from the database because their particular conformations did not allow an automatic modelling procedure. These 8 compounds were later manually inspected and re-modelled, and used as the "completely" external validation set. In total 1048 descriptors were calculated by means of the CODESSA PRO software package.

To reduce the risk of chance correlation [36], [37] and overfitting of data [38], the data set is analysed using filters to remove descriptors with either small variance or no unique information. The dataset was pre-processed column-wise and row-wise in order to eliminate constant variables, empty values and inter-correlated descriptors. In particular:

- 14 variables with standard deviation equal to zero (constant variables) were eliminated;

- eight chemicals did not have sufficient information (i.e. they had more than 80% of missing values) - these compounds contain metallic elements such as arsenic or tin, for which semi-empirical calculations cannot be performed;
- 669 variables had missing values;
- and 46 variables had an inter-correlation coefficient equal to one.

Overall, 729 variables and eight chemicals were discarded from the study leaving a dataset consisting of 274 chemicals and 319 descriptors.

The importance and relevance of pre-processing data is well know in QSAR analysis [119]. Therefore the dataset was scaled to have mean of zero and unit standard deviation.

The models generated were validated thoroughly by proper statistical procedures, including the prediction of an external validation set. In order not to bias the procedure using the external set during the tuning procedure of the model [120], such as selection of number of variables, numbers of neurons, etc., the dataset was randomly split into three separate sets as described below:

1. A training set [222 data points]: used to train the models.
2. A test set [44 data points]: used to choose the best predictive model among the different models generated.
3. A validation set [8 data points]: used to test the real predictive ability of the winning model.

Compounds involved in the study of acute aquatic toxicity are listed in Appendix A.

4.1.2 Statistical techniques

Various statistical techniques were used to extract information and to derive predictive models. All the calculations have been performed using Matlab® 7 (R14) (The MathWorks, Natick, MA).

Multiple linear regression analysis

The purpose Multiple linear regression (MLR) is to fit a linear model of the form $y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k + \varepsilon$, where y is the dependent variable (activity) and X_1, X_2,..., X_k are the independent variables (descriptors), ε is random disturbance (error), and b_0, b_1, b_2,..., b_k are the regression coefficients, which are estimated from the data finding the least square solution, i.e. regression coefficients are chosen so as to minimise the difference, i.e. error, between predicted values and actual values.

The least squares solution of the above is $b = (X^T X)^{-1} X^T y$.

Partial least squares

Partial least squares (PLS) is based on a linear transformation of the descriptors space, producing a new variable space based on a small number of orthogonal factors (latent variables), so that there is no correlation. A given number of latent variables (components) are then used as independent variables to fit a regression model. As in multiple linear regression, the main purpose of partial least squares regression is to build a linear model, $y = bX + \varepsilon$, where y is an n case by m variable response matrix, X is an n case by p variable predictor matrix, b is a p by m regression coefficient matrix, and is a noise term for the model which has the same dimensions as y. For example, suppose a data set has response variables y (in matrix form) and a large

number of predictor variables X (in matrix form), some of which are highly correlated. To establish the model, partial least squares regression produces a p by c weight matrix W for X such that $T = XW$, i.e., the columns of W are weight vectors for the X columns producing the corresponding n by c factor score matrix T. These weights are computed so that each of them maximises the covariance between responses and the corresponding factor scores. Ordinary least squares procedures for the regression of y on T are then performed to produce Q, the loadings for y (or weights for y) such that $y = TQ + \varepsilon$. Once Q is computed, we have $y = Xb + \varepsilon$, where $B = WQ$, and the prediction model is complete. One additional matrix, which is necessary for a complete description of partial least squares regression procedures, is the p by c factor loading matrix P which gives a factor model $X = TP + F$, where F is the unexplained part of the X scores. The algorithms can now be described to compute partial least squares regression.

The standard algorithm for computing partial least squares regression components (i.e., factors) is non-linear iterative partial least squares (NIPALS):

For each h=1,...,c, where $A_0 = X'y$, $M_0 = X'X$, $C_0 = I$, and c given,

1. compute q_h, the dominant eigenvector of $A_h'A_h$
2. $w_h = G_hA_hq_h$, $w_h = w_h/||w_h||$, and store w_h into W as a column
3. $p_h = M_hw_h$, $c_h = w_h'M_hw_h$, $p_h = p_h/c_h$, and store p_h into P as a column
4. $q_h = A_h'w_h/c_h$, and store q_h into Q as a column
5. $A_{h+1} = A_h - c_hp_hq_h'$ and $B_{h+1} = M_h - c_hp_hp_h'$
6. $C_{h+1} = C_h - w_hp_h'$

The factor scores matrix T is then computed as $T = XW$ and the partial least squares regression coefficients B of y on X are computed as $B = WQ$.

Back propagation neural network

Back propagation neural networks (BPNN) are among the most popular neural network architectures currently used in chemometrics. BPNNs are made up of neurons organised into layers and connected through weights. During the learning phase the weights of the BPNN are modified so that the response to a given input (the descriptors) is similar to the target (activity). Input vectors and the corresponding target vectors are used to train a network until it can approximate a function, associate input vectors with specific output vectors, or classify input vectors in an appropriate way. Standard backpropagation is a gradient descent algorithm, in which the network weights are moved along the negative of the gradient of the performance function. The term backpropagation refers to the manner in which the gradient is computed for non-linear multi-layer networks. It has been shown that a BPNN with three layers and an appropriate number of hidden neurons (neurons in the middle layer) is able to fit any function with a given accuracy [52]. A basic reference on backpropagation is the book by Rumelhart et al. [121].

For this study a three-layered network with 10-15-1 neurons using respectively tan-sigmoidal transfer function (*tansig*), tan-sigmoidal transfer function (*tansig*), linear transfer function (*pureline*) was used. The network was trained using 100 training epochs, *traingdx* learning function (a combination of adaptive learning rate with momentum training), *mse* (mean squared error) performance function, 0.01 learning rate, and a momentum constant 0.95. For details about parameters refer to the Neural Network Toolbox for Matlab® [122].

Counter-propagation neural network

Self Organizing Map (SOM) or Kohonen neural network is one of the basic types of artificial neural networks [123]. Such networks can learn to detect regularities and correlations in their input and adapt their future responses to that input accordingly. SOMs learn to recognize groups of similar input vectors in such a way that neurons physically near each other in the neuron layer respond to similar input vectors. Its architecture represents a two-dimensional grid of connected neurons, which are multi-dimensional vectors. The dimension of vectors is equal to the number of independent variables. The learning of SOM is the projection from multi-dimensional space onto two-dimensional grid (array) of neurons. The projection or learning of network runs in two-steps, the first step is the selection of the winning neuron and the second step is the self-organization of the map. In details it runs as follows. A vector, which represents an object is presented to all neurons and the algorithm selects the neuron that is most similar to it (winning neuron). In the second step the weights of the winning neuron are modified to the vector values and in the same time the neighbouring neurons are modified to become similar to it. Details and mathematical expressions are discussed in several textbooks and articles [124], [125]. After all objects are presented to the network one learning epoch is over. This procedure repeats until the weights are stabilized. As a result one obtains objects organized in two-dimensional map with layer structure, where each layer represents one component of multidimensional vector (one descriptor). The mapping is topology preserving what means that similar objects in descriptor space are located close to each other (or even on the same neuron) but it is not metric preserving. The projection from multi-dimensional space onto very limited grid of neurons caused overlapping and squeezing of information. A

map is not only a picture of original space, but also a model. In this stage of training only input variables (representation vectors) were taken into account and therefore the SOM is referred as unsupervised network. The simplest way to include the output variables (property values) is to increase the dimension of neurons and treat the input and output variables equivocally [126]. The counter propagation neural network (CPNN) implements input and output variables differently [125], [127]. The architecture of CPNN is shown in Figure 17. It has two layers the input layer, which has the same structure as in SOM and the output layer situated beneath. The difference to SOM lies in the learning strategy. The learning in the input layer is the same as in SOM, i.e., the input variables determinate the arrangement of objects. When the arrangement is set the positions of objects are projected to the output layer where the weights are modified in such a way that the weights on projected positions correspond to the output values. In addition, the weights in the neighbourhood are modified. On this way the response surface is constructed. This part of training is conducted considering the output values and therefore it is usually referred as the supervised part of training of CPNN. Similarly, the prediction runs over two steps. In the first step the object is located into input layer on the neuron with the most similar weights. In the second step, the position of that neuron is projected to the output layer, which gives the predicted output value. A detailed description of CPNN is given in [128], and [129]. CPNN has been successfully applied in a number of computational chemistry problems [130]-[134].

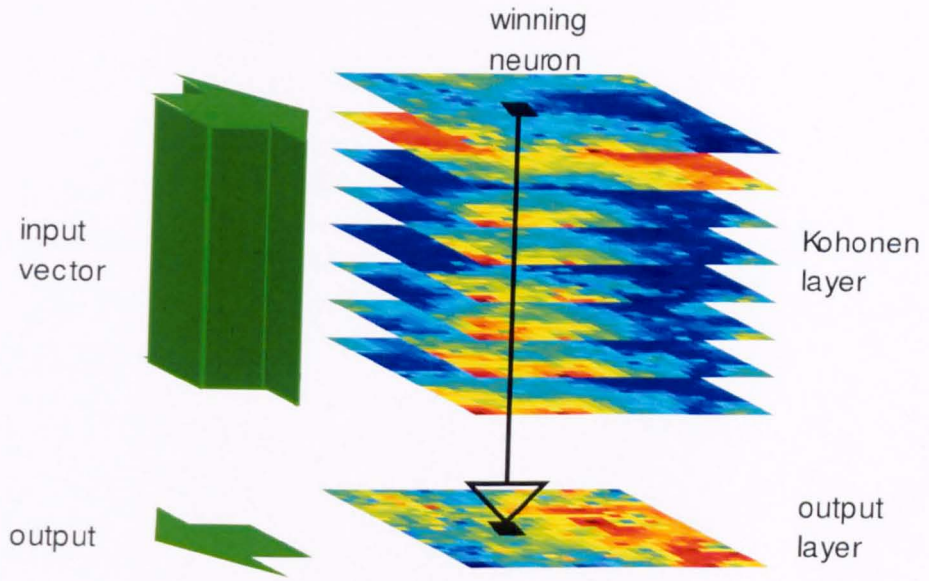


Figure 17. The architecture of CPNN.

To measure modelling performance the determination coefficient (R^2), and root mean squared error (rmse), was calculated as follows:

$$r = y - \hat{y},$$

$$rmse = \sqrt{\left(\frac{\sum r^2}{n}\right)},$$

$$sse = \sum r^2,$$

$$ssr = \sum (y - \bar{y})^2,$$

$$R^2 = 1 - \frac{sse}{ssr},$$

where y is the experimental value, is the predicted value, n is the number of data points, and is the mean of the experimental values.

4.2 RESULTS AND DISCUSSION

4.2.1 McFarland's principle

A typical QSAR model considers a term for bioavailability and a term for the reactivity of the chemicals (McFarland's principle) [135]. Generally, the

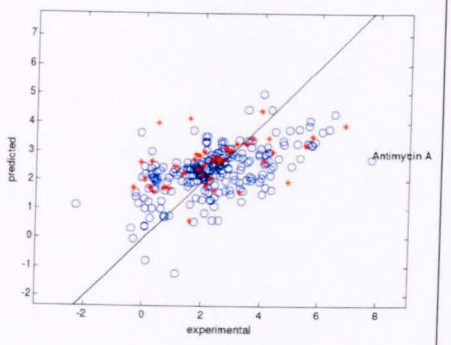
bioavailability of a chemicals to the organism is well described by the logarithm of the partition between octanol and water, i.e. LogP. Since the octanol can represent the cell membrane, LogP represents the ability of the chemical to permeate it and therefore to be available for interaction with the organism. This is especially true for the aquatic toxicity where the particular toxicological essay involve that the target species is put in a solution of a given concentration of the chemical studied.

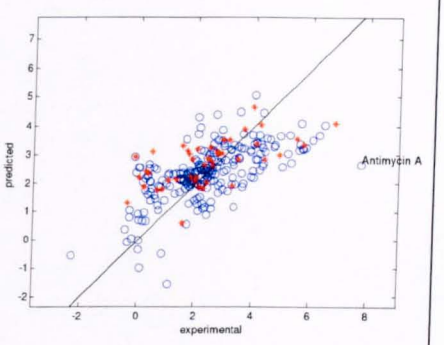
The other term is related to the chemical reactivity. In this study the energies of the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) were considered to represent it. These orbitals are called the frontier orbitals, and determine the way the molecule interacts with other species. The HOMO is the orbital that could act as an electron donor, since it is the outermost (highest energy) orbital containing electrons. Typically, the value of the highest energy molecular orbital that contains electrons is used in QSARs as it is representative of the ionisation potential. In fact, if the molecule loses an electron, it would most likely lose it from the highest energy molecular orbital, and this will change the HOMO value. The LUMO is the orbital that could act as the electron acceptor, since it is the innermost (lowest energy) orbital that has room to accept electrons. As such it represents the electron affinity in a QSAR analysis.

For this reason the first models were built-up using only LogP with either HOMO or LUMO.

Table 7. McFarland modelling results. Blue circles (o) are chemicals in the train set, red asterisks (*) are chemicals in the test set.

Tox = f(HOMO,logP)	**Tox = f(LUMO,logP)**





$Tox = 0.02HOMO + 0.40\log P + 1.24$			$Tox = -0.45LUMO + 0.37\log P + 0.95$		
	R^2	RMSE		R^2	RMSE
Train	0.32	1.282	Train	0.40	1.204
Test	0.19	1.447	Test	0.31	1.331

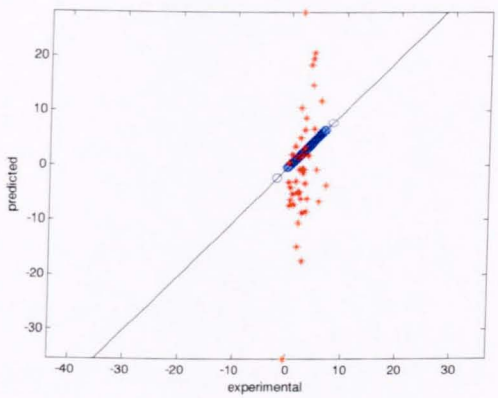
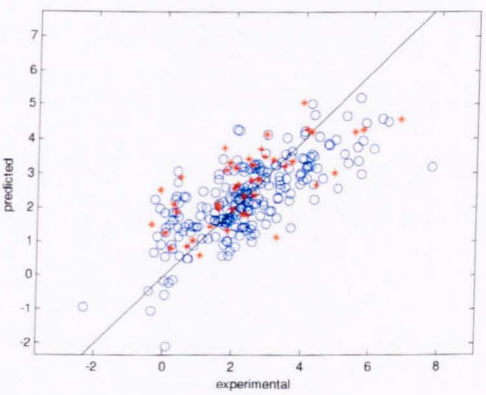
Neither the models reported in Table 7 are very predictive but do provide some encouraging indications. As in previous publications [136]-[138], the mechanism proposed is confirmed to be relevant for aquatic acute toxicity (coefficients for LogP are very similar in both models and HOMO and LUMO have, correctly, opposite signs, since they describe opposite tendencies). Moreover, a narcotic baseline effect can be recognised [139]. The only chemical not conforming to the baseline effect is Antimycin A, which is an antibiotic with a peculiar mode of action and is likely to be an outlier [140]. On the other hand, the poor performance suggests that the mechanisms needs some more parameters to be described.

4.2.2 Models of the whole dataset

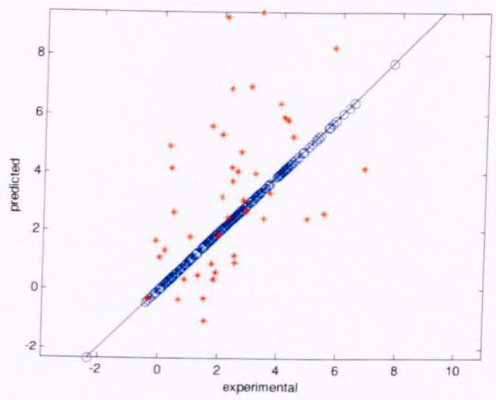
Different statistical techniques were used in order to extract relevant information from the whole dataset of 319 descriptors. Table 8 summarises the results of

some interesting analyses performed on the dataset. In particular, results obtained from MLR, PLS, BPNN are shown and commented upon here.

Table 8. Overview of the results. Blue circles (o) are chemicals in the train set, red asterisks (*) are chemicals in the test set.

a) Multiple Linear Regression		**MLR** R^2 rmse Train 1.00 0.000 Test 0.00 10.526
b) PLS (2 components)		**PLS (2 components)** R^2 rmse Train 0.53 1.063 Test 0.45 1.188

c) Backpropagation Neural Network



BPNN

	R^2	rmse
Train	1.00	0.000
Test	0.00	2.888

From the above plots, studying at each individual model, the following information can be extracted:

a. the problem is *ill-posed*. A problem is *well-posed* when a solution *exists*, is *unique* and depends *continuously* on the initial data. It is *ill-posed* when it fails to satisfy at least one of these criteria [141]. In this case, it clear that a solution might exist, but it will not be unique. In fact, the model obtained describes the training set very well (R^2 = 1.000, rmse = 0.000), but it is not able to predict the toxicity of the test set (R^2 = 0.000, rmse = 10.526). This is a solution to the problem, but it is definitely not the *true* solution we are looking for. The reasons for that, probably, are in the size of the dataset: i.e. the presence of more independent variables (319) than observations (244).

b. PLS reduces the number of variables, extracting the more relevant information from the original variables and condensing it in new orthogonal variables. The improvement of the model is clear, but the model is still not statistically significant. Some linear relationship probably exists, but the problem is highly non-linear and/or too complex to be picked up by a linear approach.

c. A non-linear model, such as the BPNN seems to be able cope the complexity of the problem, but again the presence of more variables than observation calls for variable selection.

A first general conclusion from the above is that the reliability of QSAR models have to be assessed on their predictive abilities rather than on their fitting properties which can be arbitrarily good. Moreover, it can be stated that, in this case, both relevant variable selection and non-linear modelling techniques have to be involved to build a predictive model for toxicity.

4.2.3 Selection of descriptors

A fundamental step in QSAR studies is the interpretation of the model. A good practice is to allow a mechanistic and/or biological explanation of the derived statistical model. It is intuitive that the presence of only a few variables will increase the ease of model interpretability. On the other hand, reducing the number of variables will decrease the ability of the model to explain such a complex phenomenon as toxicity.

For these reasons we choose a combination of genetic algorithm and counterpropagation neural network (GA/CPNN) to support our study. In this procedure GA explores the descriptor hyperspace for selection of variables, and CPNN derives the fitness score. For this study, 3 parallel population of 8 individual evolved through 200 generations (see section 3.3 for details), in order to optimise a 12-by-12 network, with 100 training epochs. Figure 18 shows how the determination coefficient changes in respect with the number of variables for the training (R^2_{train}) and test (R^2_{test}) sets. The analysis was conducted building models using between 1 and 20 variables selected by GA/CPNN.

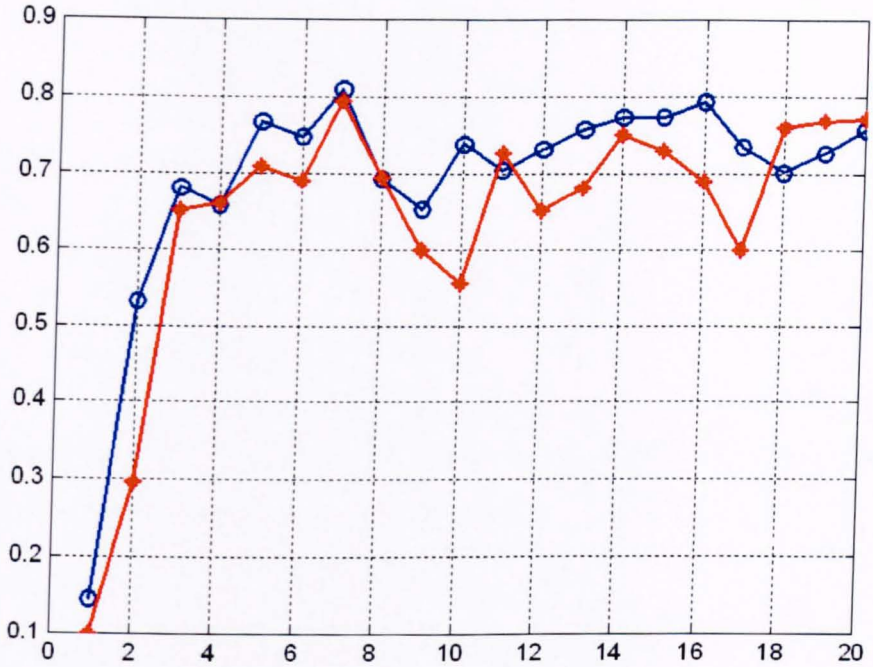
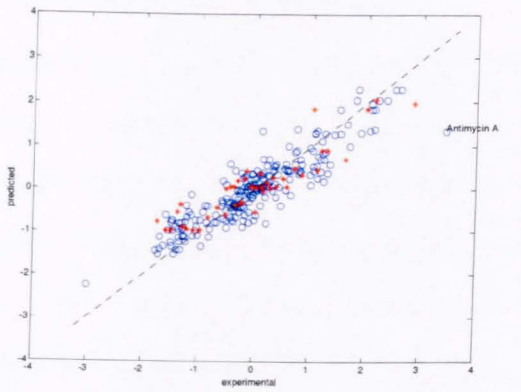


Figure 18.Performances of the model in respect with the number of variables used. Blue circles (o) are R^2_{train}, red asterisks (*) are R^2_{test}.

The model tends to a steady state after seven variables and any new variable added does not really improve the performance. The small variability around recognisable trends is due to the intrinsically stochastic nature of the modelling technique. GA searches for the "best" solution in the solution space, but since constraints are set, e.g. a maximum number of generations or a particular goal, the process may end in a local minimum.

Figure 19 shows the best model obtained by GA/CPNN using seven descriptors.



	R^2	rmse
Train	0.81	0.676
Test	0.80	0.726

Figure 19. GA/CPNN final model. Blue circles (o) are chemicals in the training set, red asterisks (*) are chemicals in the test set.

Descriptors selected are listed in the following Table 9.

Table 9. Relevant descriptors selected by GA/CPNN.

ID	Descriptor	Type[7]
1	HACA-2 (MOPAC PC)	E
2	HOMO - LUMO energy gap	QM
3	$^3\chi^v$ third order path molecular connectivity index	T
4	HA dependent HDSA-1 (Zefirov PC)	E
5	1X BETA polarisability (DIP)	QM
6	FHBCA Fractional HBSA (HBSA/TMSA) (MOPAC PC)	QM
7	LogP (KowWin1)	PC

[7] E: electrostatic; QM = quantum mechanical; T = topological; PC = physical-chemical.

4.2.4 *Interpretation of the model*

The descriptors used in the best predictive model can be divided into two main categories: penetration/solubility descriptors, which reflect the compound's abilities to form non-covalent interactions with the environment, to dissolve and persist in an aqueous or a lipid environment, or permeate the phase interfaces (i.e. LogP, hydrogen bonding descriptors, polarisability); and reactivity descriptors, which indicate the compound's abilities to interact with the surrounding molecules and form chemical bonds (i.e. the orbital gap).

Another criterion for classifying descriptors categorization is their dependence on the 3D structure. *LogP* and molecular connectivity are descriptors independent of 3D conformation, while for other descriptors a 3D (optimised) conformation must be calculated. In the OpenMolGRID system an automatic modelling procedure was applied to the structures, no conformational search was done before optimisation and the resulting 3D conformations are probably the local minimum rather than a global one. However, the promising results of the best predictive model presented show that for a heterogeneous group of compounds "any reasonable" (i.e. optimised) conformation can be used to derive 3D descriptors and construct a predictive model. This is in clear contrast to the 3D approaches used in other areas of computer-aided molecular design (especially drug design, or nanotechnologies), where the lowest energy 3D conformations for a series of conformationally (structurally) similar compounds are always sought. In any case, there are no evidences that the lowest energy conformation is actually related to the toxic mechanism.

Non-linear models such as CPNN are usually more powerful than linear ones, but are often considered "black boxes" because they do not formalise the relationship between variables and response in clear numbers or coefficients.

This raises some doubts about their use. Nevertheless, techniques to estimate the influence and weights of each variable to the model can be implemented easily. One of them is to substitute variables by a constant, for example its mean, and see how this affects the prediction (Figure 20).

In general, *LogP (KowWin1)* is a measure of the compound's lipophilicity. In aquatic species LogP describes a compound's penetration and distribution in the organism. From a physical point of view, LogP describes the entropic contributions which are important for solvation/desolvation. Numerous studies in aquatic toxicology have shown LogP to be an important descriptor frequently occurring in predictive QSAR models [136]-[138], [142].

The third order valence-corrected path molecular connectivity index ($^3\,^v$) is a topological descriptor. It encodes the presence of the hetero-atoms in the molecules with respect to their hybridisation states. Thus, in the resulting QSAR model, it can be regarded as the descriptor of molecular structure in terms of atomic connections plus the influence of hetero-atoms. The molecular index is the most important descriptor as both the determination coefficients on the training set and on the test set decrease significantly if this descriptor is kept constant (the mean value) in the descriptor test (Figure 20).

The area-weighted surface charge of hydrogen bonding acceptor atoms (HACA2), hydrogen bonding donor ability of the molecule (HDSA1), and the fractional fractional hydrogen bonding surface area divided by total molecular surface area (FHBCA) all belong to the group of surface area descriptors related to hydrogen bond formation, intermolecular interactions and compound solvation in the water environment. The apparent redundancy of hydrogen bonding-related descriptors in the model (three out of seven descriptors used in the model) is likely to be done to the presence of different subgroups of

compounds in the database (e.g. sulfonamides, carbamates) and their importance for the toxicity of these groups of compounds.

The BETA polarisability descriptor *(DIP)* reflects the molecule's properties from the point of view of polarisation induced by an external electric field and characterises the molecule as an electron acceptor. This descriptor is very important for the model, as its replacement by the constant value (the mean) significantly lowers the test determination coefficient (Figure 20).

The *HOMO-LUMO energy gap* can be regarded as a descriptor of reactivity. The HOMO-LUMO gap, i.e. the difference in energy of the highest occupied and lowest unoccupied molecular orbitals determines the compound's stability. The greater the difference, the lower the reactivity of the molecule. As it is the only "true" descriptor of the chemical reactivity in the model, it contributes substantially to the overall performance (Figure 20) in terms of both R^2_{train} and R^2_{test}.

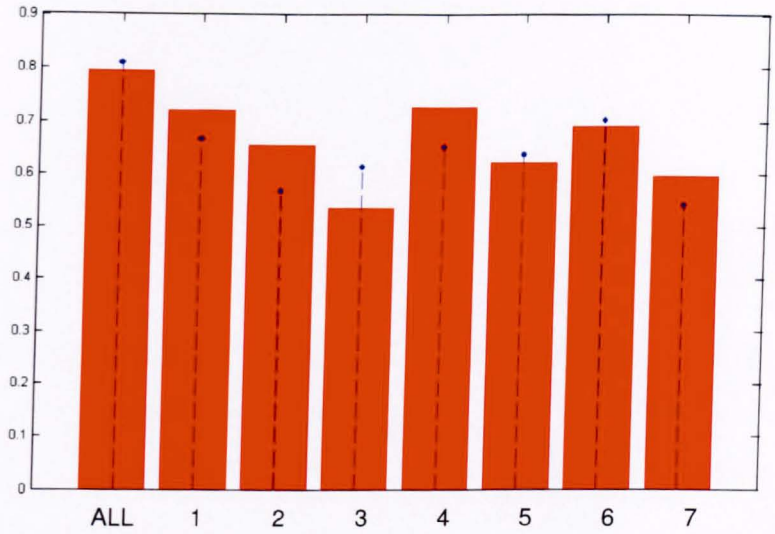


Figure 20. Descriptor analysis, showing the influence of setting variables at a constant value (mean) on the overall performances of the model. Blue stems are R^2_{train}, red areas are R^2_{test}.

4.2.5 Validation of descriptors

As already stated in the introduction to this thesis, molecular descriptors (MD) are the result of mathematical operations which transform the chemical information encoded within a symbolic representation of a molecule. Unfortunately, such a numerical representation is not unique. Indeed, each descriptor is expected to show a variability which depends strongly on the level of chemical theory behind it. For example, 2D constitutional descriptors will not change with molecular conformation. Regardless of the computational chemistry method used, the values of these descriptors are expected to match each other perfectly. On the other hand, 3D descriptors, especially quantum-mechanical ones, are much more sensitive than any other descriptors with respect to molecular structure. In fact, the use of different optimisation

procedures leads to different 3D geometries, thus to different values of 3D molecular descriptors. The key point of the current investigation is not to quantify the MD variability exactly, but to determine to what extent these values are comparable to each other. Having comparable MD values means having a QSAR model that is not dramatically dependent on the exactness of the 3D chemical structure.

In order to make this analysis, three sets of descriptors using respectively, MNDO, PM3, and AM1 methods [30] have been generated by the descriptor calculation workflow above described (paragraph 3.3), and then analysed using the following criteria:

1. Descriptor Average Standard Deviation (DASTD), defined as the mean standard deviation of each value of the *j*-th descriptor: $DASTD_j = \frac{\sum_i std(D_{i,j})}{n}$
2. Descriptor Variability Range (DVR), defined as the difference between the maximum and the minimum value of the *j*-th descriptor: $DVR_j = Max(D_j) - Min(D_j)$.
3. Descriptor Variability Percentage (DVP%), defined as: $DVP\%_j = \frac{DASTD_j}{DVR_j} \cdot 100$. This parameter indicates the average variability within the maximum range of possible values it assumes. DVP% does not depend on the absolute value of a single descriptor, providing a concrete mean to compare the variability of diverse descriptors.

Having *n* compounds and *m* descriptors, $D_{i,j}$ are the values that the *j*-th descriptor has for the *i*-th structure according to the three different parameterisations, and D_j are all the values of the *j*-th descriptor.

Table 10. Validation of the descriptors used for the acute aquatic toxicity model.

	DASTD	DVR	DVP%
HACA-2 (MOPAC PC)	0.276	10.277	2.7
HOMO – LUMO energy gap	0.295	11.342	2.6
$^{3}{}^{v}$ third order path molecular connectivity index	0.000	15.181	0.0
HA dependent HDSA-1 (Zefirov PC)	4.504	186.560	2.4
1X BETA polarizability (DIP)	131.080	6232.900	2.1
FHBCA Fractional HBSA (HBSA/TMSA) (MOPAC PC)	0.015	0.379	3.9
LogP (KowWin1)	0.000	20.57	0.0

The results above (Table 10) indicate an excellent consistency of the descriptors relevant to the QSAR model proposed, among different 3D geometries. This fact allows for a wider and simpler applicability of the model.

The definition of the applicability domain of any QSAR is still an open issue, because it raises doubts about the validity of interpolation and/or extrapolation in multidimensional spaces [143], [144]. Despite this boundaries (see Table 11) are usually useful in order to assess the chemical space of QSARs.

Table 11. Boundaries of property and relevant descriptors for the acute aquatic toxicity data sets.

	Train		Test		Validation	
	min	max	min	max	min	max
$Log_{10}(1/LC_{50})$	-2.335	7.739	-0.356	6.843	0.512	6.028
HACA-2 (MOPAC PC)	0	10.915	0	5.029	0.118	6.835

HOMO - LUMO energy gap	2.935	14.7	6.718	12.101	8.438	10.442
$^{3}{}^{v}$ third order path molecular connectivity index	0	15.181	0.643	10.08	0.274	11.306
HA dependent HDSA-1 (Zefirov PC)	0	174.15	0	129.3	5.726	107.83
BETA polarizability (DIP)	-993.29	1041.2	-681.23	244.7	-254.36	366.43
FHBCA Fractional HBSA (HBSA/TMSA) (MOPAC PC)	0	0.384	0	0.226	0	0.082
LogP (KowWin1)	-5.92	9.82	-1.1	8.39	0.62	6.38

4.2.6 Additional testing and validation of the best predictive model

The final model developed was subsequently validated using the response permutation test, also known as y-scrambling [66], [67]. This procedure involves fitting several models, in our case 100, on the same dependent variables (X block) but on a permutated response. If a strong correlation remains between the descriptors selected and the randomised response, then the significance of the proposed QSAR model is regarded as suspect.

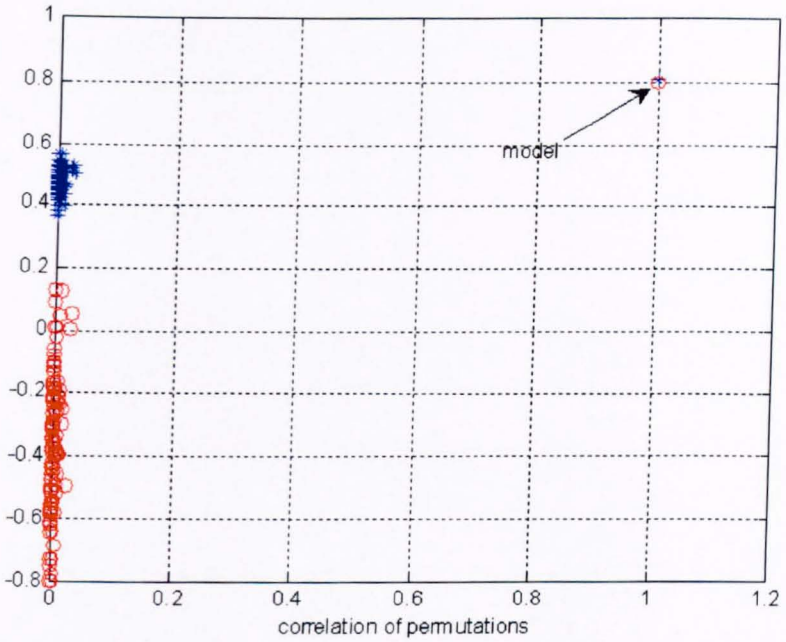


Figure 21. Response permutation testing: on the y-axis the performances of the model (R^2_{train}, as blue circles, R^2_{test}. as red asterisks), and on the x-axis the correlation between original and scrambled response.

The results of permutation testing are shown in Figure 21. The model performs much better than any of the permuted models. This is clear proof that the model is not affected by any chance correlation, and it is likely to depict a true relationship.

A further test was done to assess the reliability of the model. A sound model should be stable, and not too sensitive to noise. To simulate the influence of noisy data on the performance of the model, we added some randomness to the X block. Given the *r*-by-*c* matrix of the X block, where *r* is the number of objects (chemicals) and *c* the number of descriptors, for each *c*-th column *r* uniformly distributed random numbers in the interval (0,1) are calculated, scaled by a given percentage *n* of the standard deviation of the *c*-th descriptor and

added as noise to the column. Thus we have a new *r*-by-*c* matrix which is the original X block plus a given noise *n*. This dataset is then used to fit a model to predict the original response. For each noise *n* 50 runs were done. Results are summarized in Figure 22.

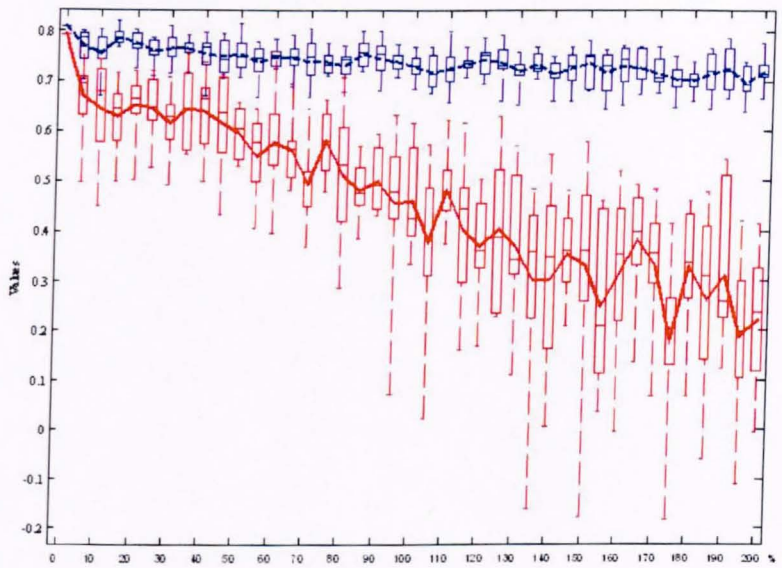


Figure 22. Sensitivity test. Ten models are fitted for each level of noise. Boxes have lines showing the lower quartile, median, and upper quartile of each level of noise. The whiskers extend from each end of the box to show the extent of the rest of the data. R^2_{train} is shown in blue (hatched), R^2_{test} in red. Marked lines show the means for each level of noise.

Figure 22 nicely illustrates a behavior common to the majority of neural networks. Observing R^2_{train} (blue lines) the model "learns" the noise as well as the "true" data. This is known as overfitting, and occurs when a learning algorithm is allowed adapting too well the training set, using for example too

many variables and/or training epochs. Hawkins defines overfitting as "the use of models or procedures that violate the principle of parsimony, ... or Occam's Razor" and gives an exhaustive description of the problem [38]. The reliability of QSAR models must therefore be assessed on their predictive ability rather than on the fitting properties which may arbitrarily be good.

In line with these general comments, we focused on the predictive power of the model, i.e. R^2_{test} (red lines). This test itself of course does not ensure that the model reflects a true relationship, but points to a reassuring behavior, stability: the model performances smoothly worsen as the noise increases, but do not present any chaotic phenomenon, or in other words there is no sensitive dependence on initial conditions. Moreover, during this test "new" objects are generated, that far being real chemicals are at least acceptable in that their descriptor values are not too distant from real values (the perturbation added is a fraction of the standard deviation). Again, the model generates reasonable prediction in relation to possible inputs.

The model was finally tested on a small set of completely external data (eight compounds), called the validation set (Figure 23). The toxicity values, LC_{50}, of these compounds ranged from 0.51 to 6.03 in a inverse logarithmic scale, and were predicted with rmse = 0.771. These results prove the robustness of the model.

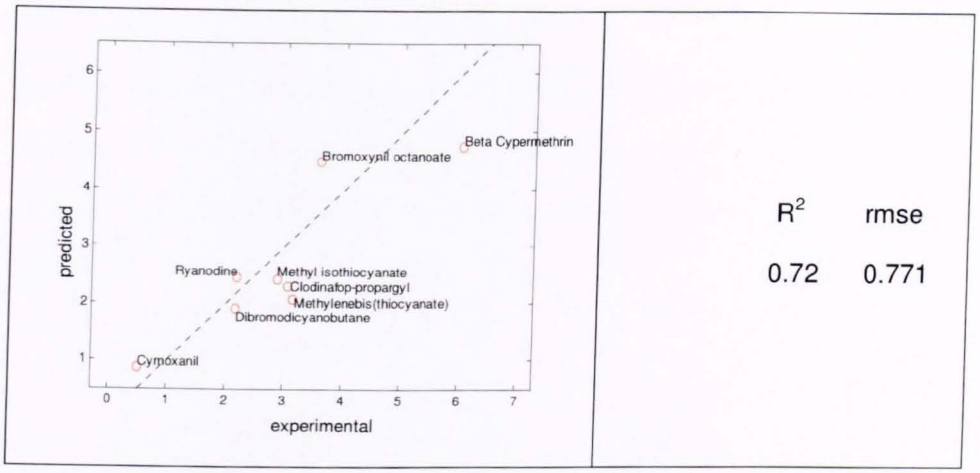


Figure 23. Predicted versus observed toxicity for the external validation set.

4.3 CONCLUSIONS

The main outcome of this study was the development of a predictive model for acute aquatic toxicity. The rigorous procedure adopted to test the QSAR model ensures its applicability and reliability to predict the toxicity of new unknown pesticides, making it particularly suitable for regulatory purposes. The mechanism emerging from analysis of the model and its descriptors is consistent with McFarland's principle for biological activity, i.e. the activity (toxicity) of a given compound is a function of the compound's abilities to penetrate (lipophilicity, hydrophilicity) and interact with biological structures (reactivity).

Generally classical linear methods can provide useful preliminary information, but can not solve complex QSAR problems. They may work on a local model, but not for a structurally diverse database like the trout LC_{50} studied in this work. Non-linear methods such as neural networks cope with the complexity of the problem but they dramatically suffer from overfitting, so a features selection is essential to reduce instrinsic variability and improve the generalizability of the

model. Among the different techniques tested, the GA/CPNN combination proved suitable for the development of ecotoxicological QSARs, because it can extract useful information hidden in the numbers and it is flexible enough to detect the non-linear relationships between molecular descriptors and biological activity.

5. AVIAN ORAL TOXICITY

This section of the thesis summarises the work done on the development of a QSAR model for the prediction of the avian oral toxicity of pesticides.

5.1 MATERIALS AND METHODS

5.1.1 Dataset

Oral toxicity LD_{50} data for the bobwhite quail (*Colinus virginianus*) were screened as described previously (paragraph 3.1). The protocol adopted provided reliable ecotoxicity values for 116 compounds.

The EU defines a classification for avian oral toxicity (Table 12) in the EC regulations 92/32EEC.

Table 12. EU class definitions of avian oral toxicity.

	LD_{50} [mg/kg]	Total number of compounds
Class 1	< 5	4
Class 2	[5 – 50[	28
Class 3	[50 – 500[	24
Class 4	≥ 500	60

Unfortunately, there are insufficient data in the most toxic class, and models derived from this distribution, probably, would not be robust enough and predictive. Because of this, the first two classes were grouped together and the distribution of activity, as indicated in Table 13, was used to develop the classification algorithms.

Table 13. Classes used for modelling, after regrouping the first two toxic classes.

	LD_{50} [mg/kg]	Total number of compounds
Class 1	< 50	32
Class 2	[50 – 500[	24
Class 3	≥ 500	60

The chemical structures of these compounds were optimised and supplemented with descriptors as described in paragraph 3.3. In total 1048 descriptors were calculated. To reduce the risk of chance correlation [36], [37] and overfitting of data [38], the data set was analysed using filters to remove descriptors with either small variance or no unique information. The dataset has been pre-processed column-wise and row-wise in order to eliminate constant variables, empty values and inter-correlated descriptors. In particular:

- 111 variables with a standard deviation equal to zero (constant variables) were eliminated;
- three chemicals did not have sufficient information (more than 80% missing values) - these compounds contain metal elements such as arsenic or tin, for which semi-empirical calculations cannot be performed;
- 574 variables had missing values;
- and 49 variables had an inter-correlation equal to one.

Therefore, 734 variables and three chemicals were discarded from the study leaving a dataset consisting of 113 chemicals and 314 descriptors.

The importance and relevance of pre-processing data is well known in QSAR analysis [119], therefore the dataset was scaled to have zero mean and unit standard deviation.

The models generated were validated thoroughly by proper statistical procedures, including the prediction of the toxicity of an external validation set. The dataset was randomly split into three separate sets as described below:

1. A training set [94 data points]: used to train the models.
2. A validation set [19 data points]: used to test the real predictive ability of the successful model.

The distribution of the pre-processed dataset is shown in Table 14.

Table 14. Distribution of the dataset used for developing the classification models.

	Number of compounds in the training set	**Number of compounds in the validation set**
Class 1	27	5
Class 2	17	5
Class 3	50	9

The compounds involved in the study of avian oral toxicity are listed in Appendix B.

5.1.2 Statistical techniques

The algorithms used were developed with Matlab® (The MathWorks, Natick, MA). The classifiers were implemented in PRTools4, a Matlab® toolbox for pattern recognition, developed at the Delft University of Technology [145], [146].

Principal component analysis

In the data mining there is typically a large number of variables in the database for it to be successful. In such situations it is highly likely that subsets of variables are highly correlated with each other. The accuracy and reliability of a classification, or prediction, model will suffer if highly correlated variables, or variables that are unrelated to the outcome of interest, are included. Superfluous variables can increase the data-collection and data-processing costs of deploying a model on a large database. The dimensionality of a model is the number of independent or input variables used by the model. One of the key steps in data mining is to find ways to reduce dimensionality without sacrificing accuracy.

Principal components analysis (PCA) is a quantitatively rigorous method for achieving the simplification of complex data matrices. The method generates a new set of variables, called principal components. Each principal component is a linear combination of the original variables. The objective of PCA is to reduce the dimensionality (number of variables) of the dataset but retain most of the original variability in the data. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

This procedure performs PCA on the selected dataset. A principal component analysis is concerned with explaining the variance covariance structure of a high dimensional random vector through a few linear combinations of the original component variables. Consider a p-dimensional random vector $\underline{X} = (X_1, X_2, ..., X_p)$. k principal components ($k < p$) of $\underline{X}$ are k (univariate) random variables $Y_1, Y_2, ..., Y_k$ which are defined by the following formulae:

$$Y_1 = l_1'\underline{X} = l_{11}X_1 + l_{12}X_2 + \ldots + l_{1p}X_p$$
$$Y_2 = l_2'\underline{X} = l_{21}X_1 + l_{22}X_2 + \ldots + l_{2p}X_p$$

$$Y_k = l_k'\underline{X} = l_{k1}X_1 + l_{k2}X_2 + \ldots + l_{kp}X_p$$

Where the coefficient vectors l_1, l_2 ,..etc are chosen such that they satisfy the following conditions:

Y_1 = Linear combination $l_1'\underline{X}$ that maximises Var($l_1'\underline{X}$) and $\| l_1 \| = 1$

Y_2 = Linear combination $l_2'\underline{X}$ that maximises Var($l_2'\underline{X}$) and $\| l_2 \| = 1$

and Cov($l_1'\underline{X}$, $l_2'\underline{X}$) =0

Y_j = Linear combination $l_j'\underline{X}$ that maximises Var($l_j'\underline{X}$) and $\| l_j \| = 1$

and Cov($l_k'\underline{X}$, $l_j'\underline{X}$) =0 for all k < j

This indicates that the principal components are those linear combinations of the original variables which maximise the variance of the linear combination and which have zero covariance (and hence zero correlation) with the previous principal components. It can be proved that there are exactly p such linear combinations. However, typically, the first few of them explain most of the variance in the original data.

Fisher linear discriminant analysis

Linear discriminant analysis (LDA) is a technique for classifying a set of observations into predefined classes. The purpose is to determine the class of an observation based on a set of variables known as predictors or input variables. The model is built based on a set of observations for which the classes are known. This set of observations is sometimes referred to as the training set. Based on the training set, the technique constructs a set of linear functions of the predictors, known as discriminant functions, such that

$$L = b_1x_1 + b_2x_2 + \ldots + b_nx_n + c,$$

where the *b*s are discriminant coefficients, the *x*'s are the input variables or predictors and *c* is a constant.

These discriminant functions are used to predict the class of a new observation with unknown class. For a *k* class problem, *k* discriminant functions are constructed. Given a new observation, all the *k* discriminant functions are evaluated and the observation is assigned to class *i* if the i^{th} discriminant function has the highest value.

K-nearest neighbours

The *k*-nearest neighbours (KNN) classification rule is a technique for non-parametric supervised pattern classification. Given the knowledge of *N* prototype patterns (vectors of dimension *D*) and their correct classification into several classes, it assigns an unclassified pattern to the class that is most heavily represented among its *k* nearest neighbours in the pattern space.

The first formulation of the above rule appears to have been made by Fix and Hodges [147] in 1951. The KNN decision rule makes no assumption on the underlying probabilistic distribution of the samples points and of their classification.

In KNN prediction, the training data set is used to predict the value of a variable of interest for each member of a target data set. Generally speaking, the algorithm is defined as follows:

1. For each row (case) in the target data set (the set to be predicted), locate the k closest members (the k nearest neighbours) of the training data set. A Euclidean distance measure is used to calculate how close each member of the training set is to the target row that is being examined.

2. Find the weighted sum of the variable of interest for the k nearest neighbours (the weights are the inverse of the distances).
3. Repeat this procedure for the remaining rows (cases) in the target set.

Support vector machine

A support vector machine (SVM) is a new and very promising classification method developed by Vapnik at al [148]. A detailed description of the theory of SVM can be found in several excellent books and tutorials [149]-[151]. Recently there has been an explosion in the number of research papers on the topic of SVMs. SVMs have been applied successfully to a number of applications ranging from particle identification, face detection, and text categorisation to engine knock detection, bioinformatics, and database marketing. The approach is systematic, reproducible, and properly motivated by statistical learning theory. Training involves optimisation of a convex cost function: there are no false local minima to complicate the learning process. SVMs are the most well-known of a class of algorithms that use the idea of kernel substitution and which are referred to as kernel methods. The general SVM and kernel methodology appears to be well-suited for data mining tasks. To understand the power and elegance of the SVM approach, one must grasp three key ideas: margins, duality and kernels.

Let us consider a binary classification task with data points x_i (i = 1,2,...,m) having corresponding labels $y_i = \pm 1$. Each data point is represented in a d-dimensional input or variable space. Let the classification function be: $f(x) = sign(w \cdot x - b)$, where w determines the orientation of a discriminant plane, and b determines the offset of the plane from the origin. In the first illustration it may be assumed that the two sets are linearly separable, i.e. a plane exists that

correctly classifies all the points in the two sets. There are a infinite number of possible separating planes that correctly classify the training data (Figure 24).

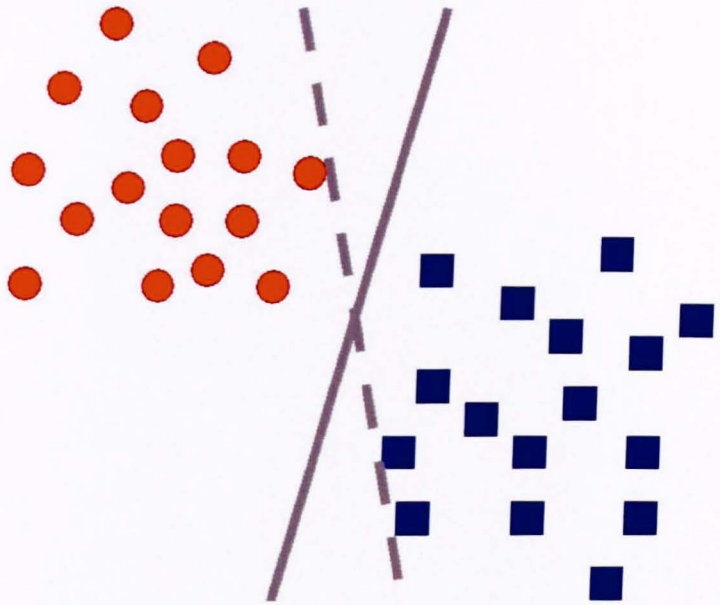


Figure 24. Linearly separable classification task with two possible discriminant planes.

Intuitively the best solution is the plane furthest from both classes (the solid line in Figure 24) because small perturbations of any point would not introduce misclassification errors, and are more likely to generalise on future data better. To construct such a plane, the convex hull of each class of training data is examined and then the closest points in the two convex hulls found (Figure 25). The convex hull of a set of points is the smallest convex set containing the points (circles labelled b and c in Figure 25).

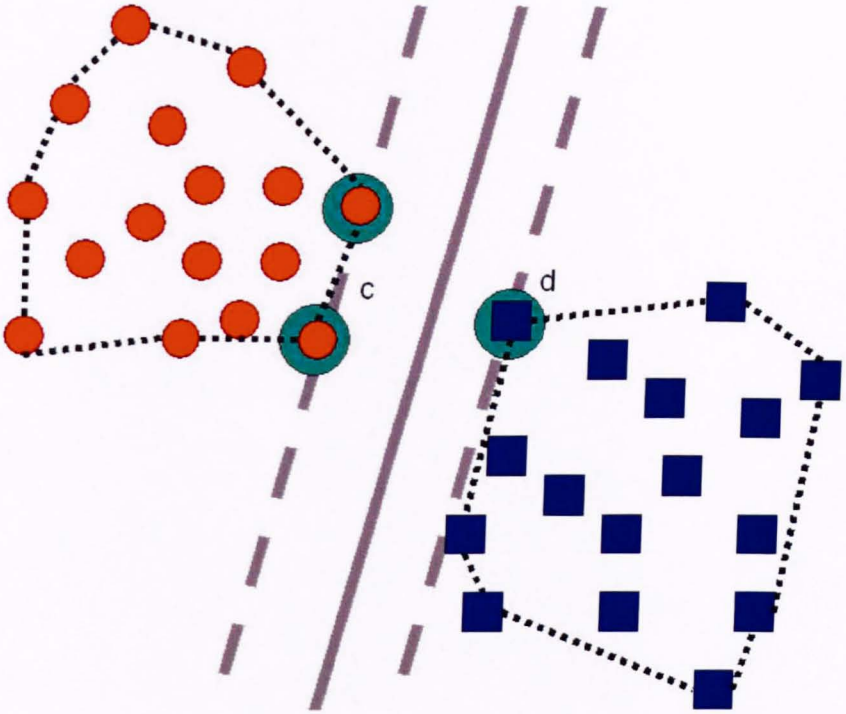


Figure 25. Best plane which bisect the closest points in the convex hull (dotted lines).

There are many existing algorithms for solving general-purpose quadratic problems (QPs) and also new approaches for exploiting the special structure of SVM problems (note that the solution depends only on the three marked circled points).

To find the plane furthest from both sets, the distance or margin between the support planes can be maximised for each class as illustrated in Figure 25. The support planes are pushed apart until they bump into a small number of data points (the support vectors) from each class. The support vectors in Figure 25 are points c and d. These same support vectors determines also the closest points in the convex hull. It is no coincidence that the solutions are identical.

This is an excellent example of the mathematical programming concept of duality.

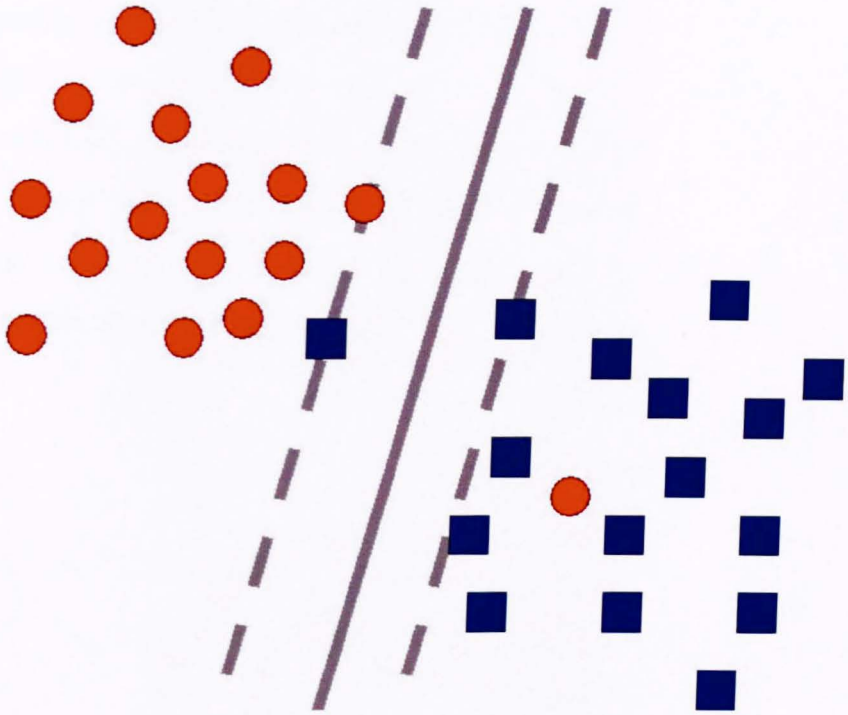


Figure 26. Selection of a plane to maximise margin and minimize error in a linearly inseparable case.

For a case that is linearly inseparable, the primal supporting plane method will fail. Since the QP task is not feasible for a linearly inseparable case, the constraints must be relaxed. Ideally no points would be to be misclassified and no points fall in the margin. However, the constraints must be relaxed to ensure that each point is on the appropriate side of the supporting plane. Any point falling on the wrong side of its supporting plane is considered to be an error. Therefore the margins should be simultaneously maximised to minimise the error (Figure 26).

By using this approach to control complexity, SVMs can construct linear classification functions with good theoretical and practical generalisation properties even in very high-dimensional attribute spaces. Robust and efficient quadratic methods exist for solving the dual formulations. But, if the linear discriminants are not appropriate for the data set, resulting in high training set errors, the SVM will not perform well. In these cases the SVM approach has to be generalised to construct highly non-linear classification functions. Consider the classification problem in Figure 27.

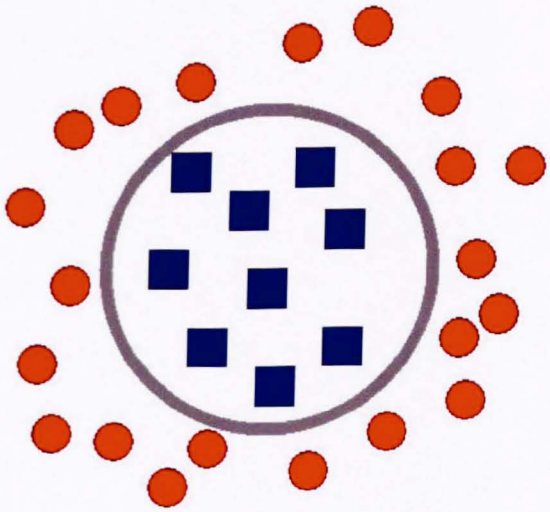


Figure 27. Example of a classification problem requiring a quadratic discriminant.

No simple linear discriminant function will work well for the data represented in Figure 27. Instead a quadratic function such as the circle pictured in Figure 27 is required. A classical method for converting a linear classification algorithm

into a non-linear classification algorithm is simply to add additional attributes to the data that are non-linear functions of the original data. Existing linear classification algorithms can be applied to the expanded dataset in feature space producing non-linear functions in the original input space. For high-dimensional datasets, this non-linear mapping method has two potential problems stemming from the fact that the dimensionality of the feature space explodes exponentially. The first problem is that of overfitting. SVMs are largely immune to this problem since they rely on margin maximisation. The second is that it is not practical to actually compute non-linear functions. SVMs get around this issue through the use of kernels. To change from a linear to a non-linear classifier, one must only substitute a kernel evaluation in the objective, instead of the original dot product. Thus by changing kernels, different and highly non-linear classifiers are obtained. No algorithm changes are required from the linear case other than substitution of a kernel evaluation for the simple do t product. All the benefits of the original linear SVM method are maintained. A highly non-linear classification function, such as a polynomial or a radial basis function machine, or a sigmoidal neural network can be trained using robust, efficient algorithms that have no problems with local minima. By using kernel substitution a linear algorithm (only capable of handling separable data) can be turned into a general non-linear algorithm.

The performance of the models were measured using the error rate (ER_{train} for training set and $ER_{validation}$ for validation set), calculated as follow: $ER = \frac{n_e}{n_t}$, where n_e is the number of misclassified object and n_t is the total number of object in the data set.

Misclassification matrices also helped in the evaluation of the performance of the models.

5.2 RESULTS AND DISCUSSION

5.2.1 Models on the whole dataset

The dataset was analysed initially by fitting the classification algorithms without any descriptor space reduction. Table 15 summarises the results obtained.

Table 15. Summary of the results for the classification models for the whole dataset.

ID	Classification algorithm	ER_{train}	$ER_{validation}$
1	LDA	0.000	0.526
2	KNN	0.287	0.579
3	SVM	0.000	0.526

It can be observed that all the algorithms have excellent fitting properties, but conversely none of them is able to generalise well on the validation set. This may again be due to the lack of an adequate number of data points with regard to the large number of variables.

5.2.2 Principal component analysis

Often, in data sets with many variables, groups of variables are related. One reason for this is that more than one variable may be measuring the same driving principle governing the behaviour of the system. In many systems there are only a few such driving forces. For these reasons PCA has been performed on the dataset in order to extract relevant information on the dataset.

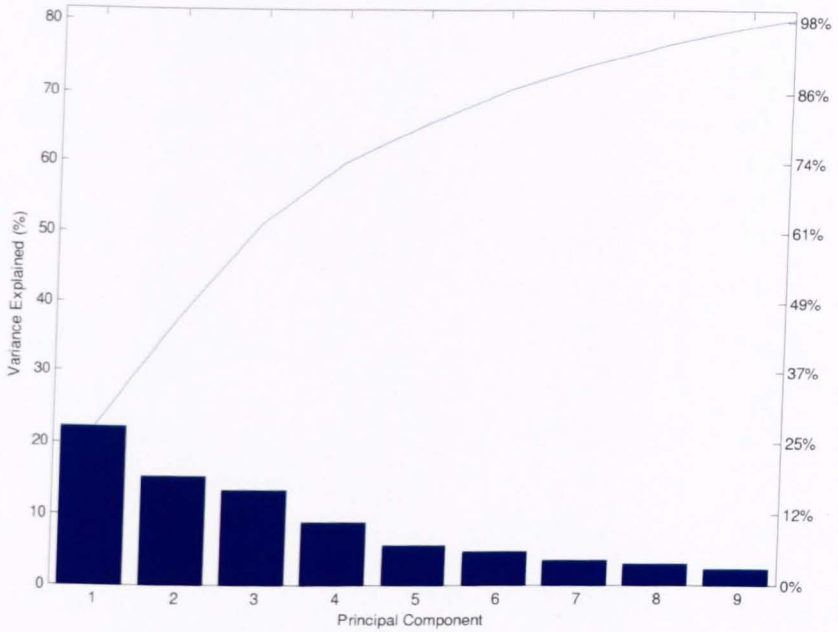


Figure 28. Variance explained by PCA for the datasets. The blue line is the cumulative variance explained.

Figure 28 shows the variance of the dataset explained by each one of the principal components extracted. It reveals that few components are not sufficient to explain the variance of the entire dataset.

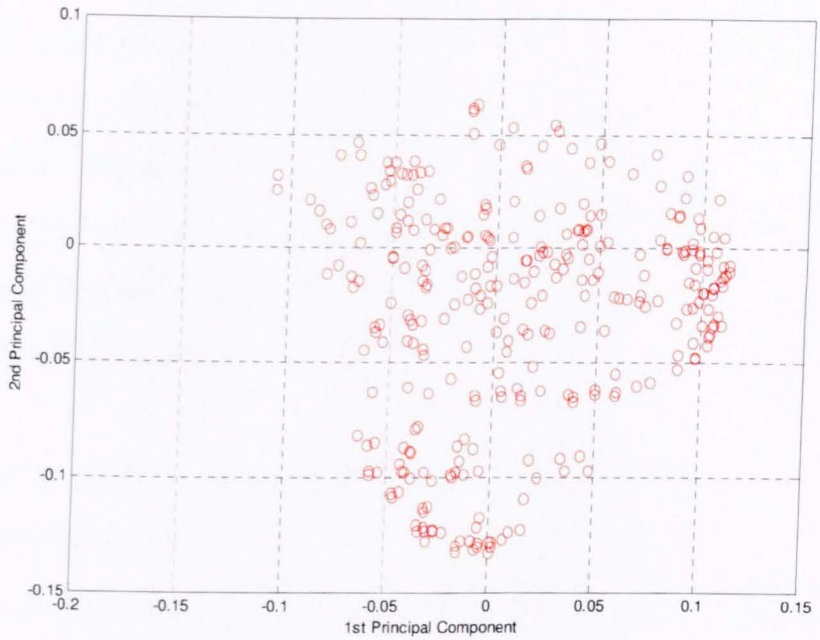


Figure 29. Loadings plot (37.68% of total variance) for 1st and 2nd principal components.

The loadings are the contribution of the descriptor vector to each of the principal components. A large positive component means that that descriptor is positively correlated with that principal component; a large negative values is negative correlation; and small values mean that the descriptor is unrelated to that principal component. Although none single descriptor stands clearly out from the loadings plot (Figure 29) in either the 1st component or the 2nd component, it is possible to identify major descriptors (absolute loading values greater than 1.1) according PCA (Table 16 and Table 17).

Table 16. Descriptors with absolute loading value of the 1st principal component greater that 0.11.

Descriptor	1st loading

Molecular volume	0.1134
TMSA Total molecular surface area (Zefirov PC)	0.1133
Molecular surface area	0.1115
Shadow plane ZX	0.1105
$^{0}{}^{v}$, zero order path molecular connectivity index	0.1104
Information content (order 0)	0.1104
Total number of atoms	0.1104

Table 16 lists the main descriptors relevant to the 1st component and contains several descriptors that refer to the general dimension of molecules.

Table 17. Descriptors with absolute loading value of the 2nd principal component greater that 0.11.

Descriptor	2nd loading
HA dependent HDCA-2 (MOPAC PC)	0.1326
HA dependent HDCA-2/SQRT(TMSA) (MOPAC PC)	0.1324
HA dependent HDCA-1 (MOPAC PC)	0.1311
HA dependent HDCA-1 (Zefirov PC)	0.1301
HA dependent HDCA-2 (Zefirov PC)	0.1299
HA dependent HDCA-2/SQRT(TMSA) (Zefirov PC)	0.1296
HA dependent HDSA-1 (MOPAC PC)	0.1286
HA dependent HDSA-2 (MOPAC PC)	0.1280
HA dependent HDSA-1 (Zefirov PC)	0.1278
HA dependent HDSA-2/SQRT(TMSA) (MOPAC PC)	0.1276
HA dependent HDCA-2/TMSA (MOPAC PC)	0.1273

HA dependent HDSA-2/SQRT(TMSA) (Zefirov PC)	0.1272
HA dependent HDSA-2 (Zefirov PC)	0.1271
min(#HA, #HD) (MOPAC PC)	0.1239
HA dependent HDSA-2/TMSA (Zefirov PC)	0.1235
HA dependent HDSA-2/TMSA (MOPAC PC)	0.1233
HA dependent HDCA-1/TMSA (MOPAC PC)	0.1230
HA dependent HDCA-2/TMSA (Zefirov PC)	0.1229
min(#HA, #HD) (Zefirov PC)	0.1229
HA dependent HDSA-1/TMSA (Zefirov PC)	0.1224
HA dependent HDSA-1/TMSA (MOPAC PC)	0.1218
count of H-acceptor sites (MOPAC PC)	0.1213
HA dependent HDCA-1/TMSA (Zefirov PC)	0.1202
count of H-acceptor sites (Zefirov PC)	0.1173
HASA-2 (MOPAC PC)	0.1146
HACA-2 (Zefirov PC)	0.1132
HACA-2 (MOPAC PC)	0.1124

The main descriptors relevant to the 2nd component (Table 17) encode the ability of molecules to form hydrogen bonds. It is clear that these characteristics, dimension and hydrogen-bonding interaction, play relevant roles in the dataset.

The score are the data points in the new coordinate system defined by the first two principal components. The score plot (Figure 30) is the plot of the score of the first two principal components and is useful to determine the distribution of compounds in the chemical space defined by the descriptors.

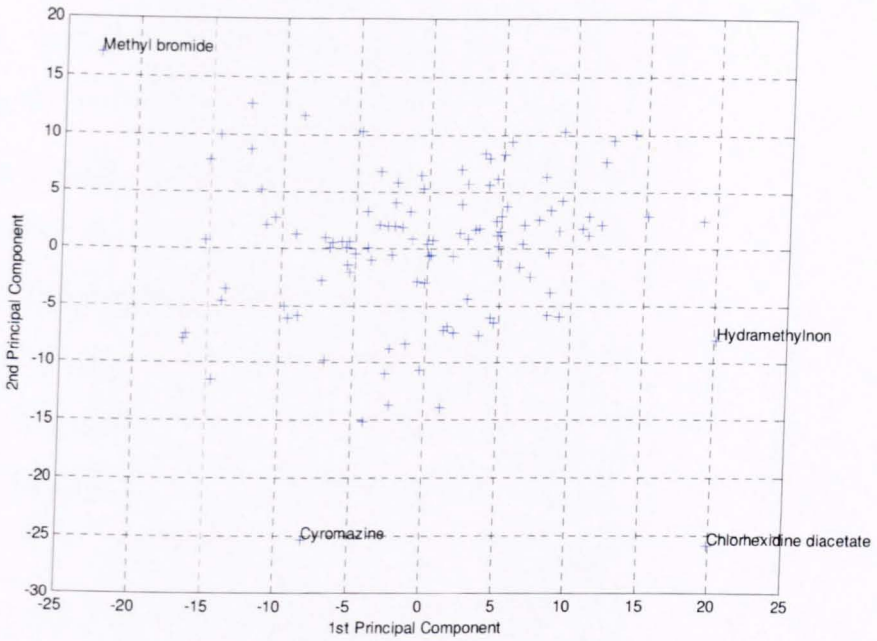


Figure 30. Score plot (37.68% of total variance) of 1st and 2nd principal components.

Using the first two components, that retrieve the maximum amount of the variance (37.68%), it is possible to observe how distant or different are chemicals in the dataset. Chemicals that are far from each other on this plot are also chemically very diverse. In this case, it can be observed that the compounds cover the chemical space very well.

Hotelling's T^2 test is a statistical measure of the multivariate distance of each observation from the centre of the data set. This is an analytical method to find the most extreme points in the data. Compounds at the borders -methyl bromide, hydramethylnon, chlorhexidine diacetate, and cyromazine- have the largest Hotelling's T^2 value, and the most diverse chemicals.

Classifiers were trained on the descriptor space reduced by PCA. 95% of the total variance, i.e. the first nine principal components (Figure 28), was extracted. Table 18 summarises the results of the model obtained.

Table 18. Classification models in the descriptor space reduced by PCA (95% of total variance).

ID	Classification algorithm	ER_{train}	$ER_{validation}$
4	PCA/LDA	0.223	0.474
5	PCA/KNN	0.287	0.526
6	PCA/SVM	0.191	0.474

PCA helped to improve the generalisability of classification, but again, none of the models reflect a true relationship between descriptors and toxicity. This is indicated by the performance of the external validation set, which is far worse than the performance for the training set.

5.2.3 Selection of descriptors

A GA (see section 3.3 for details) was used to select variables from the dataset. The objective function was the error rate (ER) of the model obtained from fitting a SVM classifier implementing the Radial Basis function kernel defined below:

$$\exp\left(-\frac{\|u-v\|^2}{2\sigma}\right)$$

where u, and v are any two points of the dataset and is the standard deviation.

The best classification model (Table 19) was obtained using the nine variables listed in Table 20.

Table 19. Classifiers fitted on variables selected by GA with different fitness functions (validated on test set).

ID	Classification algorithm	# variables	$ER_{training}$	$ER_{validation}$
7	GA/SVM	9	0.021	0.158

Table 20. Descriptors selected by GA.

ID	Name	Type[8]
1	PPSA-3 Atomic charge weighted PPSA (MOPAC PC)	QM
2	Number of S atoms	C
3	Positively Charged Part of Partial Charged Surface Area (MOPAC PC)	E
4	$^{2}\chi^{v}$, second order path molecular connectivity index	T
5	Min e-e repulsion for atom C	QM
6	Min net atomic charge for atom C	QM
7	Molecular weight	C
8	Bonding Information content (order 2)	T
9	HA dependent HDCA-2/TMSA (Zefirov PC)	E

The atomic charge weighted partial positive surface area (PPSA3) is obtained by the summation of products of the individual atomic partial charges and the atomic-accessible surface areas.

The positively charged part of partial charged surface area (PPSA1) is defined as the sum of the positively charged solvent-accessible atomic surface areas.

[8] QM = quantum mechanical; C = constitutional; E: electrostatic; T = topological.

These descriptors are expected to encode the features responsible for polar interactions between molecules.

Sulphur atoms are often present in pesticides compounds because they may play a role in the biochemical interactions between the molecule and the biological system, i.e. specific mechanisms of action.

The valence connectivity indices were suggested to account for the presence of hetero-atoms and the hybridisation of atoms in the molecule. The definition of those descriptors proceeds from the atomic valence connectivity for the *i*-th atom in the molecular skeleton. The second order relates to atoms of two contiguous bond fragments.

The electron-electron repulsion energy describes the electron repulsion-driven process and may be related to conformational changes (rotation, inversion) or atomic reactivity in the molecule. As related to a given atom, it may specify the site of a particular chemical activity or conformational change in the molecule.

The net atomic charge for an atom describes the deficiency or sufficiency of the electron population of the atom in a molecule and reproduces the electrostatic potential around a given molecule.

Molecular weight is the sum of the weights of the atoms of which it is made. It can encode the macroscopic dimension of molecules, and can serve to introduce a molar quantity into the modelling and the mechanism of action.

The bonding information content is a topological descriptor based on information theory, which according to Shannon's statistical information theory [152] can be viewed as a measure of the mean quantity of information contained in each structural element.

The HA dependent HDCA-2/TMSA is an electrostatic descriptor defined to account for the possible hydrogen-bonding interactions between the molecules.

It is calculated as the sum of the area weighted surface charge of hydrogen-bonding donor atoms in the molecule normalised by the total molecular surface area.

5.2.4 Analysis of the model

A detailed analysis of the model was conducted studying the misclassification matrix of both training (Table 21) and validation sets (Table 22).

Table 21. Misclassification matrix for the training set.

	Predicted classes			
True classes	**1**	**2**	**3**	**Totals**
1	27	0	0	27
2	0	15	2[9]	17
3	0	0	50	50
Totals	27	15	52	94

Only two compounds are misclassified in the training set (Imidacloprid and Cyproconazole). They both belong to the second class (LD_{50} = 50-500 mg/kg) but are predicted in the third (LD_{50} > 500 mg/kg), and therefore predicted less toxic than actually they are. Similarly, in the validation set, the three misclassified compounds are assigned to the less toxic class (third class) although they belong to the first (4-Aminopyridine) or the second class (Chinomethionat, DBNPA).

[9] Imidacloprid, Cyproconazole.

Table 22. Misclassification matrix for the validation set.

	Predicted classes			
True classes	**1**	**2**	**3**	**Totals**
1	4	0	1[10]	5
2	0	3	2[11]	5
3	0	0	9	9
Totals	4	3	12	19

From a toxicological prospective, the existence of false positives (chemicals that are predicted less toxic than they actually are) are much more dangerous than false negatives (chemicals that are predicted more toxic than they actually are). For this reason the misclassified compounds were analysed in depth.

PCA was used to study the chemical space of the variables selected. The first two components of PCA already explain about 55% of the total variance of the data set (Figure 31). Thus, the space described by the first and second principal components can, to a certain extent, be considered representative of the complete chemical space.

[10] 4-Aminopyridine.

[11] Chinomethionat, DBNPA.

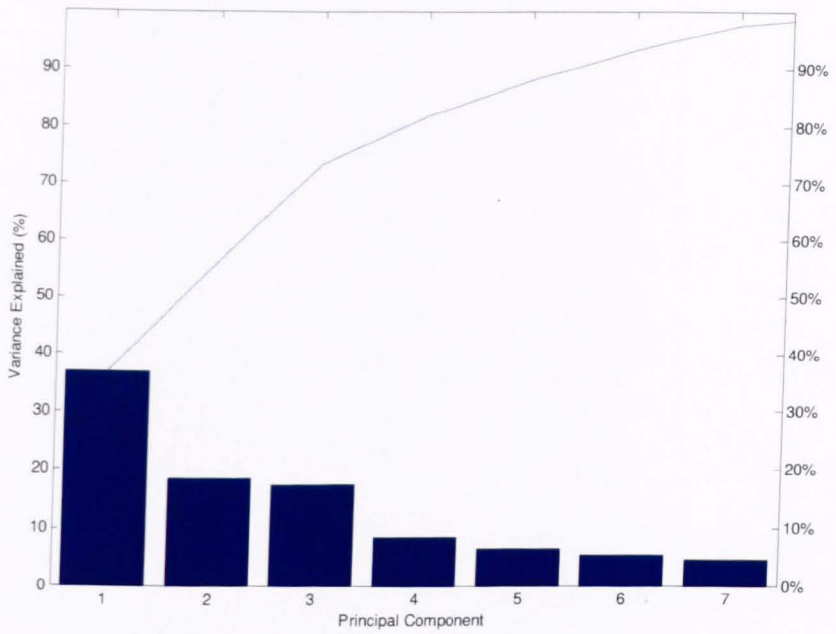


Figure 31. Variance explained by PCA for the selected variables. The blue line is the cumulative variance explained.

It is worth noting from the loadings plot (Figure 32) that descriptors encoding the presence of hetero atoms and molecular dimensions, i.e. the number of Sulphur atoms, $^{2}\chi^{v}$, and molecular weight (respectively number 2, 4, and 7 in Figure 32) are clustered in the upper-left side of the plot, and are mutually correlated. It is also clear from PCA that molecules with heavy hetero-atoms have low values of the Positively Charged Part of Partial Charged Surface Area (MOPAC PC) and HA dependent HDCA-2/TMSA (Zefirov PC) descriptors (respectively number 3, and 9 in Figure 32). Thus, molecules with high values for these descriptors are not willing to form polar interactions and hydrogen-bonds, probably because of small atomic-accessible surface areas.

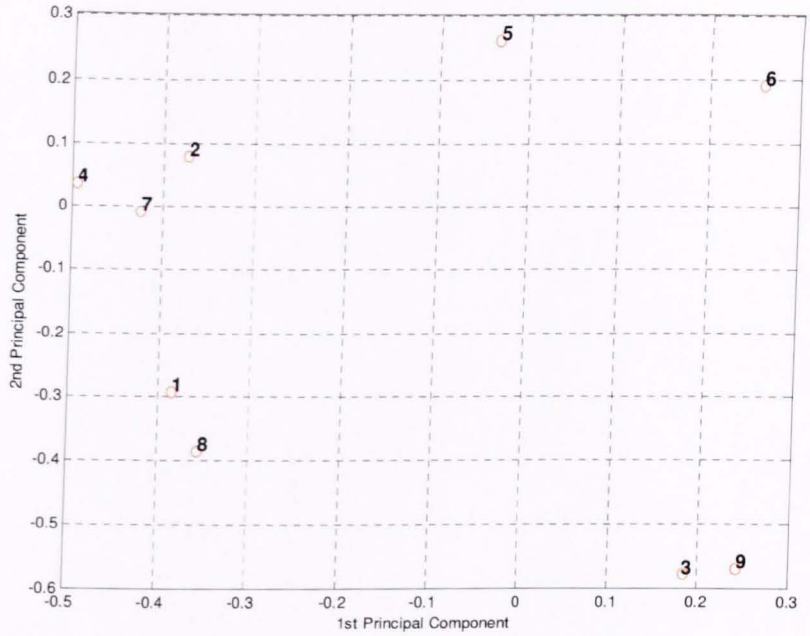


Figure 32. Loadings plot of the selected variables (55.92% of total variance) for 1st and 2nd principal components. Variables are labelled by their ID (Table 20).

The scores plot (Figure 33) reveals the relative position of the misclassified compounds in the chemical space of the variables selected.

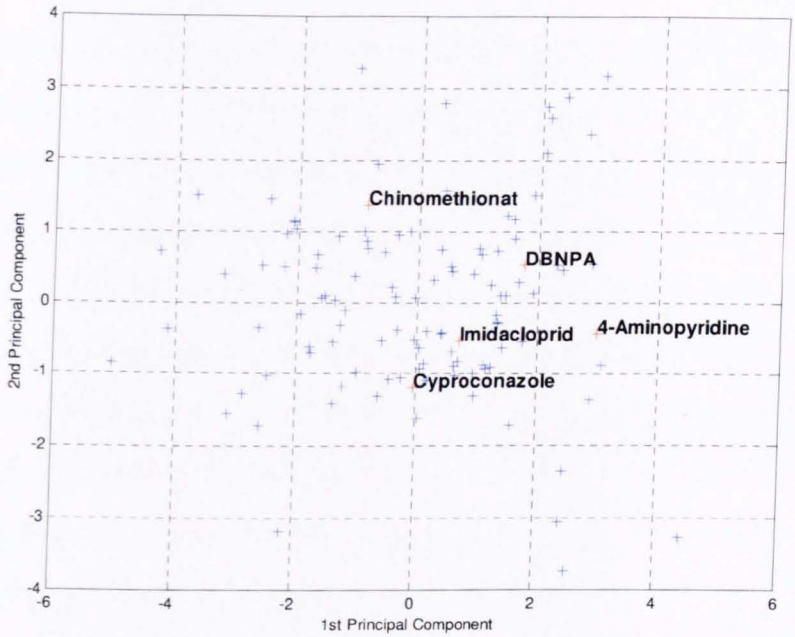


Figure 33. Scores plot of the selected variables (55.92% of total variance) of 1st and 2nd principal components. Misclassified compounds from the model are marked in red.

From Figure 33 it is intuitive that all the five misclassified compounds are not too structurally diverse, as compared to the rest of the dataset. The PCA places them in areas well populated and not marginal, with the exception of 4-Aminoprydine. The error in the prediction could therefore be caused by experimental error or the fact that these compounds may act with specific mechanisms of action that are not described by the variables selected. 4-Aminoprydine, which is also the most poorly predicted compound, is a small molecule with high tendency to form hydrogen bonds and there is not a similar molecule in the training set (see Appendix B).

5.2.5 Validation of descriptors

As previously (Paragraph 4.2.5), the variability of molecular descriptors (MD) with regard to the method used to obtain the optimised 3D structure was studied. Of course, some descriptors, e.g. 2D constitutional descriptors, are not affected by molecular conformation. On the other hand, 3D descriptors, especially quantum-mechanical ones, are much more sensitive than any other descriptors with respect to molecular structure and geometry. In fact, the use of different optimisation procedures leads to different 3D geometries, and thus to different values for 3D molecular descriptors. The key point of the current investigation is not to quantify MD variability exactly, but to determine to what extent these values are comparable to each other. Having comparable MD values means having a QSAR model that is not dramatically dependent on the exactness of the 3D chemical structure.

In order to make this analysis, three sets of descriptors using, respectively, MNDO, PM3, and AM1 methods [30] have been generated by the descriptor calculation workflow above described (paragraph 3.3), and then analysed using the following criteria:

1. Descriptor Average Standard Deviation (DASTD), defined as the mean standard deviation of each value of the *j*-th descriptor: $DASTD_j = \frac{\sum_i std(D_{i,j})}{n}$
2. Descriptor Variability Range (DVR), defined as the difference between the maximum and the minimum value of the *j*-th descriptor: $DVR_j = Max(D_j) - Min(D_j)$.
3. Descriptor Variability Percentage (DVP%), defined as: $DVP\%_j = \frac{DASTD_j}{DVR_j} \cdot 100$. This parameter indicates the average variability

within the maximum range of possible values it assumes. DVP% does not depend on the absolute value of a single descriptor, providing a concrete mean to compare the variability of diverse descriptors.

Having n compounds and m descriptors, $D_{i,j}$ are the values that the j-th descriptor has for the i-th structure according to the three different parameterisations, and D_j are all the values of the j-th descriptor.

Table 23. Validation of the descriptors used for the avian oral toxicity.

	DASTD	DVR	DVP%
PPSA-3 Atomic charge weighted PPSA (MOPAC PC)	12.456	76.205	16.35
Number of S atoms	0	4	0
Positively Charged Part of Partial Charged Surface Area (MOPAC PC)	0.0015	0.0443	3.39
$^{2}\chi^{v}$, second order path molecular connectivity index	0.000	15.436	0.00
Min e-e repulsion for atom C	3.488	46.252	7.54
Min net atomic charge for atom C	0.0915	1.6626	5.50
Molecular weight	0.00	488.53	0.00
Bonding Information content (order 2)	0.000	54.948	0.00
HA dependent HDCA-2/TMSA (Zefirov PC)	0.0005	0.0105	5.07

As opposed to the modelling of aquatic acute toxicity (Table 10), the descriptors used to modelling avian oral toxicity are generally dependent on the 3D conformation of the chemical (Table 23). Apart from 2D constitutional and topological MD, the descriptors have a significant variability (DVP%) associated

with the 3D structure of chemicals. This means that the model is sensitive to the accuracy of the 3D structures.

The definition of the applicability domain of any QSAR is still an open issue, because it raises doubts about the validity of interpolation and/or extrapolation in multidimensional spaces [143], [144]. Despite this boundaries (see Table 24) are usually useful in order to assess the chemical space of QSARs.

Table 24. Boundaries of descriptors selected for the avian oral toxicity data sets.

	Train		Validation	
	min	max	min	max
PPSA-3 Atomic charge weighted PPSA (MOPAC PC)	9.7026	88.561	15.125	79.587
Number of S atoms	0	4	0	3
Positively Charged Part of Partial Charged Surface Area (MOPAC PC)	0.0018698	0.03127	0.010205	0.03417 7
$^{2}\chi_{v}$, second order path molecular connectivity index	0	15.933	0.88172	13.401
Min e-e repulsion for atom C	54.825	81.428	57.457	73.213
Min net atomic charge for atom C	-0.8627	0.4317	-0.878	-0.1115
Molecular weight	94.939	577.93	91.109	397.5
Bonding Information content (order 2)	3.4274	43.837	6.4211	32.348
HA dependent HDCA-2/TMSA (Zefirov PC)	0	0.0037566	0	0.0037429

5.3 CONCLUSIONS

The aim of this study was the QSAR analysis of avian oral toxicity. Because of the dimensions and characteristics of the data set, classification techniques were preferred, and the data were organised into three classes of decreasing toxicity. Several statistical techniques have been adopted in order to build up a predictive model for avian oral toxicity, including linear and non-linear classifiers, PCA and the use of a genetic algorithm for variable selection.

The best classification model was chosen on the basis of the performances on a validation set of 19 data points, and was obtained fitting a support vector machine using 94 data points and nine variables selected by genetic algorithms. The model allowed for a mechanistic estimation of the toxicological action. In fact, several descriptors selected for the final classification model encode for the interaction of the pesticides with other molecules, the presence of hetero-atoms, e.g. sulphur atoms, is correlated with the toxicity, and the pool of descriptor selected is generally dependent from the 3D conformation of the structures. These suggest that, in the case of avian oral toxicity, pesticides probably exert their toxic action through the interaction with some macromolecule and/or protein of the biological system. In fact, the interaction of pesticides with steroid receptors, microsomal enzymes and cholinesterase activity is confirmed by *in vivo* tests [76]-[81].

6. DISCUSSION AND CONCLUSIONS

As described in the introduction, any QSAR is based on three fundamental sciences: biology, chemistry, and mathematics. The experience gained during this thesis allowed for the evaluation of the adequacy and maturity of these in QSAR studies.

Statistical techniques and computational power placed at the disposal of today's researchers, allow for reliable and high throughput data analysis. These techniques span regression and classification approaches, linear and non-linear methods, and are able to cope well with the increasing complexity of QSAR studies. The power of these techniques often requires thorough validation of the results, but robust methods exist and are available to the researcher. It is clear that the assessment of the applicability domain of the findings has to be derived for the chemical space used for the validation, but at the moment rigorous and reliable methods are still missing. The current research in this area is focused on the evaluation of the density of the data sets used.

The chemical information involved in QSAR studies can be considered exhaustive. Thousands of different descriptors are calculated by researchers and it is foreseeable that they can cover and explain the great part of the chemistry hidden in the structures. Of course, new specific descriptors might be useful but the limitations of QSARs are probably not to be ascribed to chemical information. This is especially true when modern non-linear techniques are used, because possible lack of terms derived from the manipulation and/or interaction of existing descriptors is overcome by these mathematical approaches. A different reasoning has to be performed for the determination of 3D conformers. Numerous 3D structures has been studied carefully by X-ray

crystallography and collected in databases. In addition, today's powerful theoretical methods can give precise approximations of actual minimum energy conformers. Despite this, it should be remembered that, the active compound is not always the most energetically stable, and often-neglected particular conditions of the biological system, such as pH or temperature, can dramatically influence 3D structures. Moreover, chemicals encounter a number of transformation when they enter a biological system, and the actual toxic action observed and measured may be caused by a metabolite rather, than the original parent compound. Thus, a next step, that can greatly improve QSAR studies, is taking into account in the analysis the presence of metabolites together with a deeper analysis of the real conditions of the biological system, especially when interactions with proteins are expected.

However, the main limitation to QSAR is probably due to the quality and abundance of the biological data and/or information. QSAR models remain, of course, a mathematical approximation of a phenomenon that is not governed by mathematical rules, thus it will always be affected by errors. A second, non-trivial problem, of biological data is their intrinsic variability. Common ecotoxicological endpoints are macroscopic measures of the conditions of a population, e.g. LC_{50} and LD_{50}, and often large variability is observed during the experimental tests. The external variables and the state variables of the ecological system are not usually observed or not observable and therefore the information that could help in the evaluation of such variability is lost. Nevertheless, under rigorous conditions, QSAR studies can provide useful and valuable information, on the condition that there are enough "numbers" to work on.

In light of the above reasoning, the QSAR obtained for the prediction of acute aquatic toxicity can be trusted to reflect a true relationship between the descriptors selected and toxicity values. The main outcome of this study was the development of a predictive model for acute aquatic toxicity. The descriptive statistics of the model are very encouraging and the rigorous procedure adopted to test the QSAR model ensures its applicability and reliability to predict the toxicity of new unknown pesticides, making it particularly suitable for regulatory purposes. Thanks to its small mean error, this predictive model is apt for use to prioritise *in vivo* tests, and for screening potential chemicals. The mechanism emerging from analysis of the model and its descriptors is consistent with McFarland's principle for biological activity, i.e. the activity (toxicity) of a given compound is a function of the compound's abilities to penetrate (lipophilicity, hydrophilicity) and interact with biological structures (reactivity and/or receptor binding).

In the case of the avian oral toxicity, the lack of ecotoxicological data did not allow for a solid and rigorous validation of the model. Predictions from the models are promising, but the model is better used to study the mechanism of action rather than for the prediction of toxicity. In fact, the findings can be useful to understand something about the mechanisms of the avian oral toxicity, for which experimental studies and evidence are still largely missing. The descriptors involved in the model suggest that the presence of hydrogen-bonding donor atoms in the molecule tend to reduce the avian oral toxicity, whereas small molecules with hetero-atoms are likely to be more toxic to the bobwhite quail.

Finally, two general comments were also derived from this study: i) non-linear methods are generally more powerful than linear ones in QSAR studies

because they are able to better adapt to the complexity of biological system; ii) variable selection is a fundamental step in the development of predictive models, and genetic algorithms were shown to be particularly apt to this purpose.

7. BIBLIOGRAPHY

[1] Carson, R.L. 1962. Silent Spring. Riverside Press, Cambridge, MA, USA.

[2] Pimetel, D. 1995. Amounts of pesticides reaching target pests: environmental impacts and ethics. *J. Agric. Envrion. Ethics*, 8, pp. 47-84.

[3] WHO-UNEP. 1989. Public Health impact of pesticides used in agriculture. World Health Organization-United Nations Environment Programme, Geneva, Switzerland.

[4] Levine, R. 1991. Recognized and possible effects of pesticides in humans. In: W.J. Hayes and E.R. Laws (Editors), Handbook of Pesticide Toxicology, Academic Press, San Diego, CA, pp. 275-360.

[5] OEDC Guidelines for Testing Chemicals, Method 203, Fish Acute Toxicity Test; Paris, 1984. Adopted July 17, 1992.

[6] OPP 72-1. Acute Toxicity Test for Freshwater Fish (Pesticide Assessment Guidelines, Subvision E-Hazard Evaluation; Wildlife and Aquatic Organisms), EPA report 540/09-82-024, 1982.

[7] OPPTS 850.1075. Ecological Effects Test Guidelines, Fish Acute Toxicity Test, Freshwater and Marine, EPA 712-C-96-118, 1996. (see http://www.epa.gov/opptsfrs/OPPTS_Harmonized/850_Ecological_Effects_Test_Guidelines/Drafts/850-1075.pdf).

[8] OPP 72–6 Aquatic Organism Accumulation Tests (Pesticide Assessment Guidelines, Subdivision E—Hazard Evaluation: Wildlife and Aquatic Organisms) EPA report 540/09–82–025 (1982).

[9] 40 CFR 797.2175 Avian Acute Oral Toxicity Test.

[10] OPP 71–1 Avian Single-Dose LD50 Test (Pesticide Assessment Guidelines, Subdivision E—Hazard Evaluation; Wildlife and Aquatic Organisms) EPA report 540/09-82-024, 1982.

[11] US EPA OPPTS 850.2100 Avian Acute Oral Toxicity Test (Ecological Effects Test Guidelines). 1996.

[12] Purcell, W.P.; Bass, G.E.; Clayton, J.M. Strategy of Drug Design. A molecular Guide to Biological Activity. Wiley, New York, 1973.

[13] Tute, M.S. 1990. History and Objectives of Quantitative Drug Design, in: Ramsden, C.A. (Ed.), *Quantitative Drug Design*, Volume 4 of: Hansch, C., Sammes, P.G. and Taylor, J.B. (Eds.). Comprehensive Medicinal Chemistry. The Rational Design, Mechanistic Study & Therapeutic Application of Chemical Compounds, Pergamon Press, Oxford, pp. 1-31.

[14] Rekker, R.F. 1992. The history of Drug Research: From Overton to Hansch. *Quant. Struct.-Act. Relat.* 11, pp. 195-199.

[15] van de Waterbeemd, H. 1992. The History of Drug Research: From Hansch to the Present, *Quant. Struct.-Act. Relat.* 11, pp. 200-204.

[16] Kubinyi, H. 2002. From Narcosis to Hyperspace: The History of QSAR. *Quant. Struct.-Act. Relat.* 21, pp. 348-356.

[17] Cros, A.F.A. 1863. PhD Thesis, University of Strasbourg; cited from: S. Borman 1990. New QSAR Techniques Eyed for Environmental Assessments, *Chem. & Eng. News*, February 19, pp. 20-23.

[18] Crum Brown, A.; Fraser, T. R. 1868. On the connection between chemical constitution and physiologic action. Part 1. On the physiological action of salts of the ammonium bases, derived from strychnia, brucia, thebia, codeia, morphia and nicotia, *Trans. Roy. Soc.* Edinburgh 25, pp. 151-203.

[19] Hansch, C.; Fujita, T. 1964. ρ–σ–π Analysis. A method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.*, 86, pp.1616-1626.

[20] Free, S.M. Jr., and Wilson, J.W. 1964. A Mathematical Contribution to Structure Activity Studies, *J. Med. Chem.*, 7, pp. 395-399.

[21] Schultz, T.W.; Cronin, M.T.D. 2003. Essential and desirable characteristics of ecotoxicity quantitative structure-activity relationships. *Environ. Toxicol. Chem.*, 22. pp. 599-607.

[22] Schultz, T.W.; Cronin, M.T.D.; Walker J.D.; Aptula, A.O. 2003. Quantitative structure–activity relationships (QSARs) in toxicology: a historical perspective. *J. Mol. Struct.: THEOCHEM*, 622, pp. 1-22.

[23] Jaworska, J.S.; Comber, M.; Auer, C.; Van Leeuwen, C.J. 2003. Summary of a Workshop on Regulatory Acceptance of (Q)SARs for Human Health and Environmental Endpoints. *Environ. Health Persp.*, 111, pp. 1358-1360.

[24] Eriksson, L.; Jaworska, J.S.; Worth, A.P.; Cronin, M.T.D.; McDowell, R.M.; Gramatica, P. 2003. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based on QSARs. *Environ. Health Persp.*, 111, pp. 1361-1375.

[25] Cronin, M.T.D.; Walker, J.D.; Jaworska, J.S.; Comber, M.H.I.; Watts, C.D.; Worth, A.P. 2003. Use of QSARs in International Decision-Making Frameworks to Predict Ecologic Effects and Environmental Fate of Chemical Substances. *Environ. Health Persp.*, 111, pp. 1376-1390.

[26] Cronin, M.T.D.; Jaworska, J.S.; Walker, J.D.; Comber, M.H.I.; Watts, C.D.; Worth, A.P. 2003. Use of QSARs in International Decision-Making

Frameworks to Predict Health Effects of Chemical Substances. *Environ. Health Persp.*, 111, pp. 1391-1401.

[27] Unger, S.H.; Hansch, C. 1973. On Model Building in Structure-Activity Relationships. A Reexamination of Adrenergic Blocking Activity of β-Halo-β-arylalkylamines. *J. Med. Chem.* 16, pp. 745-749.

[28] OECD series on testing and assessment, Number 49, THE REPORT FROM THE EXPERT GROUP ON (QUANTITATIVE) STRUCTURE-ACTIVITY RELATIONSHIPS [(Q)SARs] ON THE PRINCIPLES FOR THE VALIDATION OF (Q)SARs, 2nd Meeting of the ad hoc Expert Group on QSARs, OECD Headquarters, 20-21 September, 2004.

[29] Egan, W. J.; Morgan, S. 1998. L. Outlier detection in multivariate analytical chemical data *Anal. Chem. 70*, pp. 2372-2379.

[30] Cramer C.J. Essentials of Computational Chemistry: Theories and Models, ISBN: 0-471-48552-7, 562 pages.

[31] Todeschini, R.; Consonni, V. 2000. Handbook of Molecular Descriptors, Volume 11 of: Mannhold, R.; Kubinyi, H.; Timmerman, H. (Eds.), *Methods and Principles in Medicinal Chemistry*, Wiley-VCH, Weinheim.

[32] Goodford, P.J. 1985. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* 28(7), pp. 849-857.

[33] Cramer, R. D.; Patterson, D. E.; Bunce, J. D. 1988. Comparative Molecular Field Analysis (CoMFA).I. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.*, 110, pp.5959-5967.

[34] Klebe, G.; Abraham, U.; Mietzner, T. 1994. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Potency. *J. Med. Chem.*, 37, pp. 4130-4146.

[35] Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. 2000. Grid-Independent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.*, 43, pp. 3233-3243.

[36] Topliss, J.G.; Costello R.J. 1972. Chance Correlation in Structure-Activity Studies Using Multiple Regression Analysis. *J. Med. Chem.*, *15*, 10, pp. 1066-1068.

[37] Topliss, J.G.; Edwards, R.P. 1979. Chance Factors in Studies of Quantitative Structure-Activity Relationships. *J. Med. Chem.*, *22*, 10, pp. 1238-1244.

[38] Hawkins, D.M. 2004. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci*, *44*, pp. 1-12.

[39] Sutter, J.M.; Kalivas, J.H. 1993. Comparison of Forward Selection, Backward Elimination, and Generalized Simulated Annealing for Variable Selection. *Microchem. J.*, **1993,** *47*, pp. 60-66.

[40] Xu, L.; Zhang, W.-J. 2001. Comparison of different methods for variable selection. *Anal. Chim. Acta*, *446*, pp. 477-483.

[41] Mazzatorta, P.; M. Vracko, M.; Benfenati, E. 2003. ANVAS: Artificial neural variables adaptation system for descriptor selection. *J. Comput. Aid. Mol. Des. 17*, pp. 335-346.

[42] Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. 1993. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct. Act. Relat.*, *12*, pp. 9-20.

[43] Burden, F.R.; Ford, M.G.; Whitley, D.C.; Winkler, D.A. 2000. Use of automatic relevance determination in QSAR studies using Bayesian neural networks. *J. Chem. Inf. Comput. Sci.* 40, pp. 1423-30.

[44] Tetko, I.V.; Villa, A.E.; Livingstone, D.J. 1996. Neural network studies. 2. Variable selection. *J. Chem. Inf. Comput. Sci.*, 36, pp. 794-803.

[45] Castellano, G.; Fanelli, A. M. 2000. Variable selection using neural-network models. *Neurocomputing, 31*, pp. 1-13.

[46] Chatterjee, S.; Price, B. 1977. Regression Analysis by Example; Wiley: New York.

[47] Despagne, F.; Massart, D.-L. 1998. Variable selection for neural networks in multivariate calibration. *Chem. Intell. Lab. Sys., 40*, pp. 145-163.

[48] Höskuldsson, A. 2001. Variable and subset selection in PLS regression. *Chem. Intell. Lab. Syst., 55*, pp. 23-38.

[49] Jackson, J. E. 1991. *A User's Guide to Principal Components*, John Wiley and Sons, Inc., 1991, p. 592.

[50] Wold, H., 1985. Partial Least Squares in Samuel Kotz and Norman L. Johnson, eds., *Encyclopedia of Statistical Sciences, Vol.* 6, New York: Wiley, pp. 581-591.

[51] Wold S.; Eriksson, L. 1995. Statistical validation of QSAR results. In: van de Waterbeemd H, ed. *Chemometric Methods in Molecular Design New* York: VCH Publishers, Inc., pp. 309-318.

[52] Funahashi, K.-I. 1989. On the approximate realization of continuous mappings by neural networks, *Neural Networks*, 2, pp. 183-192.

[53] I. Kövesdi, M.F. 1999. Dominguez-Rodriguez, and L. Ôrfi. Application of neural networks in structure-activity relationships. *Med. Res. Rev.* 19, pp. 249-269.

[54] Weiss, S.M.; Kulikowski, C.A. 1991. Computer Systems That Learn, Morgan Kaufmann.

[55] Efron, B.; Tibshirani, R.J. 1993. An Introduction to the Bootstrap, London: Chapman & Hall.

[56] Hjorth, J.S.U. 1994. Computer Intensive Statistical Methods Validation, Model Selection, and Bootstrap, London: Chapman & Hall.

[57] Efron, B. 1982. The Jackknife, the Bootstrap and Other Resampling Planes. Philadelphia: Society of Industrial and Applied Mathematics.

[58] Shao, J. and Tu, D. 1995. The Jackknife and Bootstrap, New York: Springer-Verlag.

[59] Kubinyi, H.; Hamprecht, F.A.; Mietzner, T. 1998. Three-Dimensional Quantitative Similarity-Activity Relationships (3D QsiAR) From SEAL Similarity Matrices. *J. Med. Chem.* 41, pp. 2553-2564.

[60] Golbraikh, A.; Tropsha. A. 2002. Beware of q^2!. *J. Mol. Graph. Model., 20,* pp. 269-276.

[61] Tropsha, A.; Grammatica, P.; Vijay K. Gombar. 2003. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.*, 22, pp. 69-77.

[62] Hawkins, M.D.; Basak, S.C.; Mills, D. 2003. Assessing Model Fit by Cross-Validation. *J. Chem. Inf. Comput. Sci, 43*, pp. 579-586.

[63] Hawkins, D.M. 2004. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.*, 44, pp. 1-12.

[64] Efron, B. 1983. Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. of the American Statistical Association*, 78, pp. 316-331.

[65] Wehrens, R.; Putter, H.; Buydens, L.M.C. 2000. The bootstrap: A tutorial. *Chemometr. Intell. Lab. Sys.*, 54, pp. 35-52.

[66] Van der Voet, H. 1994. Comparing the predictive accuracy of models using a simple randomization test. *Chemometr. Intell. Lab. Sys.*, 25, pp. 313-323.

[67] Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. 2001. Multi- and Megavariate Data Analysis – Principles and Applications. Umea, Sweden: Umetrics AB.

[68] M.M. Babín M.M.; Tarazona, J.V. 2005. In vitro toxicity of selected pesticides on RTG-2 and RTL-W1 fish cell lines. *Environ. Pollut.*, 135, 2, pp. 267-274.

[69] Ferrari, A.; Anguiano, O.L.; Soleño, J.; Venturino A.; Pechen de D'Angelo, A.M. 2004. Different susceptibility of two aquatic vertebrates (Oncorhynchus mykiss and Bufo arenarum) to azinphos methyl and carbaryl. *Comp. Biochem. Phys. C*, 139, 4, pp. 239-243.

[70] Jarrard, H.E.; Delaney, K.R.; Kennedy, C.J. 2004. Impacts of carbamate pesticides on olfactory neurophysiology and cholinesterase activity in coho salmon (Oncorhynchus kisutch). *Aquat. Toxicol.*, 69, 2, pp. 133-148.

[71] Ruiz-Leal M.; George, S. 2004. An in vitro procedure for evaluation of early stage oxidative stress in an established fish cell line applied to investigation of PHAH and pesticide toxicity. *Mar. Environ. Res.*, 58, 2-5, pp. 631-635.

[72] Mager, P.P. 1982. Structure-neurotoxicity relationships applied to organophosphorus pesticides. *Toxicol. Lett.*, 11, 1-2, pp. 67-71.

[73] Mager, P.P. 1997. The Role of Diagnostic Statistics in Medicinal Chemistry. *Med. Res. Rev.*, 17, pp. 505-522.

[74] Devillers, J.; Flatin, J. 2000. A general QSAR model for predicting the acute toxicity of pesticides to Oncorhynchis mykiss. *SAR and QSAR in Environmental Research*, 11, pp. 25-43.

[75] Devillers, J. 2001. A general QSAR model for predicting the acute toxicity of pesticides to Lepomis macrochirus. *SAR QSAR Environ. Res.*, 11, pp. 397-411.

[76] Ottinger M.A.; Wu, J.; Hazelton, J.; Abdelnabi, M.; Thompson, N.; Quinn, M.; Donoghue, D.; Schenck, F.; Ruscio, M.; Beavers, J.; Jaber, M. 2005. Assessing the consequences of the pesticide methoxychlor: neuroendocrine and behavioral measures as indicators of biological impact of an estrogenic environmental chemical. *Brain Res. Bull.*, 65, 3, pp. 199-209.

[77] Eroschenko, V.P.; Amstislavsky, S.Y.; Schwabel, H.; Ingermann, R.L. 2002. Altered behaviors in male mice, male quail, and salamander larvae following early exposures to the estrogenic pesticide methoxychlor. *Neurotoxicol. Teratol.*, 24, 1, pp. 29-36.

[78] Ottinger, M. A.; Abdelnabi, M. A.; Henry, P.; McGary, S.; Thompson, N.; Wu, J. M. 2001. Neuroendocrine and Behavioral Implications of Endocrine Disrupting Chemicals in Quail. *Hormon. Behav.*, 40, 2, pp. 234-247.

[79] Stanley, P.I.; Bunyan, P.J.; Rees, W.D.; Swindon D.M.; Westlake, G.E. 1978. Pesticide-induced changes in hepatic microsomal enzyme systems: Further studies on the effects of 1,1,-di(p-chlorophenyl)-2-chloroethylene (DDMU) in the Japanese Quail. *Chem.-Biol. Interact.*, 21, 2-3, pp. 203-213.

[80] Hill, E.F. 1989. Divergent effects of postmortem ambient temperature on organophosphorus- and carbamate-inhibited brain cholinesterase activity in birds. *Pestic. Biochem. Phys.*, 33, 3, pp. 264-275.

[81] Fleming, W.J.; Grue, C.E. 1981. Recovery of cholinesterase activity in five avian species exposed to dicrotophos, and organophosphorus pesticide. *Pestic. Biochem. Phys.*, 16, 2, pp. 129-135.

[82] Roncaglioni, A.; Benfenati, E.; Boriani, E.; Clook, M. 2004. A Protocol to Select High Quality Datasets of Ecotoxicity Values for Pesticides. *J. Environ. Sci. Health, B* 39, 4, pp. 641-650.

[83] Hansch, C.; Leo, A. 1995. Exploring QSAR, Fundamentals and Applications in Chemistry and Biology, ACS, Washington, DC, Chapters 6 & 11.

[84] McKinney, J.D.; Richard, A.; Waller, C.; Newman, M.C.; Gerberick, F. 2000. The practice of structure activity relationships (SAR) in toxicology. *Toxicol. Sci.*, 56, pp. 8-17.

[85] Katritzky, A R; Petrukhin, R; Tatham, D; Basak, S; Benfenati, E; Karelson, M; Maran, U. 2001. Interpretation of quantitative structure-property and - activity relationships. *J. Chem. Inf. Comp. Sci.*, 41, pp. 679-685.

[86] Walker, J.D.; Schultz, T.W. Structure-activity relationships for predicting ecological effects of chemicals. In Hoffman D., Rattner B.A., Burton G. Jr., Cairns J. Jr., eds., *Handbook of Ecotoxicology*, 2nd ed. CRC. Boca Raton, FL, USA, pp. 893-910.

[87] Romberg, M. 2002. The UNICORE Grid Infrastructure. *Scientific Programming Special Issue on Grid Computing*, 10, pp. 149-158.

[88] Moss, L.; Adelman, S. 2000. Data Warehousing Methodology. *J. Data Warehousing*, 5, pp. 23-31.

[89] Karelson, M. 2000. Molecular Descriptors in QSAR/QSPR. John Wiley & Sons, New York.

[90] Katritzky, A.R.; Gordeeva, E.V.; Shcherbukhin, V.V.; Zefirov, N.S. 1993. Rapid Conversion of Molecular Graphs to 3D Representation using the MOLGEO Program. *J. Chem. Inf. Comput. Sci.*, 33, pp. 102-111.

[91] Stewart, J.J. 1990. MOPAC: a semiempirical molecular orbital program. *J. Comput. Aid. Mol. Des.*, 4, pp 1-45.

[92] http://www.codessa-pro.com

[93] Mazzatorta, P.; Benfenati, E.; Schuller, B.; Romberg, M.; McCourt, D.; Dubitzky, W.; Sild, S.; Karelson, M.; Papp, A.; Bágyi, I.; Darvas, F. 2004. OpenMolGRID: Molecular Science and Engineering in a Grid Context. In *Proceedings of PDPTA 2004, The 2004 International Conference on Parallel and Distributed Processing Techniques and Applications*, June 21-2 2004, Las Vegas, Nevada, USA.

[94] http://www.openmolgrid.org/

[95] Meylan, W.M.; Howard, P.H. 1995. Atom/fragment contribution method for estimating octanol-water partition coefficients. *J. Pharm. Sci.*, 84, pp. 83-92.

[96] Burden, F.R.; Ford, M.G.; Whitley, D.C.; Winkler, D.A. 2000. Use of automatic relevance determination in QSAR studies using Bayesian neural networks. *J. Chem. Inf. Comput. Sci.*, 40, pp. 1423-30.

[97] Chatterjee, S.; Price, B. 1997. *Regression Analysis by Example*; Wiley: New York.

[98] Despagne, F.; Massart, D.-L. 1998. Variable selection for neural networks in multivariate calibration. *Chem. Intell. Lab. Sys., 40*, pp. 145-163.

[99] Höskuldsson, A. 2001. Variable and subset selection in PLS regression. *Chem. Intell. Lab. Syst.*, *55*, pp. 23-38.

[100] Lindgren, F.; Geladi, P.; Wold, S. 1994. Kernel-Based Pls Regression Cross-Validation and Application to Spectral Data. *J. Chemometr.*, *8*, pp. 377-389.

[101] Sutter, J.M.; Kalivas, J.H. 1993. Comparison of Forward Selection, Backward Elimination, and Generalized Simulated Annealing for Variable Selection. *Microchem. J.*, *47*, pp. 60-66.

[102] Ros, F.; Pintore, M.; Chrétien, J. R. 2002. Molecular descriptor selection combining genetic algorithms and fuzzy logic: application to database mining procedures. *Chemom. Intell. Lab. Syst.*, *63*, pp. 15-26.

[103] Xu, L.; Zhang, W.-J.; 2001. Comparison of different methods for variable selection. *Anal. Chim. Acta*, *446*, pp. 477-483.

[104] Goldberg, D. E. 1989. Genetic Algorithms in Search, Optimization & Machine Learning; Addison-Wesley: New York.

[105] Hasegawa, K.; Miyashita, Y.; Funatsu, K. 1997. GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists. *J. Chem. Inf. Comput. Sci.*, *37*, pp. 306-310.

[106] Hibbert, D.B. 1993. Generation and display of chemical structures by genetic algorithms. *Chemom. Intell. Lab. Syst.*, *20*, pp. 35-43.

[107] Holland, J. 1975. Adaptation in Natural and Artificial Systems; The University of Michigan Press: Ann Arbor.

[108] Kinnear, K. E. 1994. Advances in Genetic Programming; MIT Press: Cambridge, MA.

[109] Baker, J. E. 1987. Reducing bias and inefficiency in the selection algorithm, *Proc ICGA 2*, pp. 14-21.

[110] Chipperfield, A. J.; Fleming, P. J. 1995. The MATLAB Genetic Algorithm Toolbox, from *IEE Colloqium on Applied Control Techniques Using MATLAB*, Digest No. 1995/014.

[111] Chipperfield, A. J.; Fleming, P. J.; Fonseca, C. 1994. M. Genetic Algorithm Tools for Control System Engineering. In *Proc. Adaptive Computing in Engineering Design and Control*, Plymouth Engineering Design Center, 21-22 September, pp. 128-133.

[112] http://www.shef.ac.uk/uni/projects/gaipp/ga-toolbox/.

[113] Booker, L. 1987. Improving search in genetic algorithms. In *Genetic Algorithms and Simulated Annealing*; L. Davis (Ed.); Morgan Kaufmann Publishers, pp 61-73.

[114] Gold, L. S.; Slone, T. H.; Manley, N. B.; Backman Garfinkel, G.; Hudes, E. S.; Rohrbach, L.; Ames, B. N. 1991. The Carcinogenic Potency Database: Analysis of 4000 Chronic Animal Cancer Experiments Published in the General Literature and by the U.S. National Cancer Institute/National Toxicology Program. *Environ. Health Perspect.*, 96, pp. 11-15.

[115] Gini, G.; Lorenzini, M.; Benfenati, E.; Grasso, P.; Bruschi, M. 1999. Predictive Carcinogenicity: A Model for Aromatic Compounds, with Nitrogen-Containing Substituents, Based on Molecular Descriptors Using an Artificial Neural Network. *J. Chem. Inf. Comput. Sci.*, 39, pp. 1076-1080.

[116] Debnath, A. K.; Debnath, G.; Shusterman, A. J.; Hansch, J. 1992. A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test: 1. Mutagenicity of aromatic and heteroaromatic amines in *Salmonella typhimurium* TA98 and TA100. *Environ. Mol. Mutagen.*, 19, pp. 37-52.

[117] Basak, S. C.; Mills, D. 2001. Prediction of mutagenicity utilizing a hierarchical QSAR approach. *SAR QSAR in Environ. Res.*, 12, pp. 481-496.

[118] Lemke, F.; M Iler, J.-A.; Benfenati, E. 2004. Modelling and Prediction of Toxicity of Environmental Pollutants. In *Knowledge Exploration in Life Science Informatics*; J.A. Lòpez, E. Benfenati and W. Dubitzky, Springer, pp. 221-234.

[119] Mazzatorta, P.; Benfenati, E.; Neagu, C.D.; Gini, G. 2002. The importance of scaling in data mining for toxicity prediction. *J. Chem. Inf. Comput. Sci.*, 42, pp. 1250-1255.

[120] Mazzatorta, P.; Benfenati, E.; Neagu, C.D.; and Gini G. 2003. Tuning neural and fuzzy-neural networks for toxicity modeling. *J. Chem. Inf. Comput. Sci.*, 43, pp. 513-518.

[121] Rumelhart, D.E.; Hinton, G.E.; Williams R.J. 1986. Learning internal representations by error propagation. In *Parallel Data Processing*; D. Rumelhart and J. McClelland, editors. Vol.1, Chapter 8, the M.I.T. Press, Cambridge, MA, pp. 318-362.

[122] http://www.mathworks.com/access/helpdesk/help/toolbox/nnet/

[123] Kohonen, T. 1987. Self-Organization and Associative Memory, 2nd Edition, Berlin: Springer-Verlag.

[124] Leardi, R. 2003. Nature-inspired methods in chemometrics: Genetic algorithms and artificial neural networks, Elsevier: Amsterdam.

[125] Zupan, J.; Gasteiger, J. 1999. Neural networks in chemistry and drug design, Wiley-VCH: Weinheim.

[126] Kawakami, J.; Hoshi, K.; Ishiyama, A.; Miyagishima, S.; Sato, K. 2004. Application of a Self-Organizing Map to Quantitative Structure-Activity

Relationship Analysis of Carboquinone and Benzodiazepine *Chem. Pharm. Bull.*, *52*, pp. 751-755.

[127] Hecht-Neilson, R. 1987. Counter propagation Networks. *Appl. Optics.*, 26, pp. 4979-4984.

[128] Dayhof, J. 1990. In *Neural Network Architectures, An Introduction*; Van Nostrand Reinhold: New York, p 192.

[129] Zupan, J.; Novic, M.; Gasteiger, J. 1995. Neural networks with counter-propagation learning strategy used for modelling. *Chemometr. Intell. Lab.* 27, 2, pp. 175-187.

[130] Mazzatorta, P.; Vra ko, M.; Jezierska, A. 2002. Toxicity Map: application of counterpropagation neural network in the investigation of aquatic acute toxicity. In *Proceedings of SKD 2002, Slovenski Kemijski Dnevi 2002*, September 26-27 2002: Maribor, Slovene, pp. 329-332.

[131] Mazzatorta P.; Vra ko M.; Jezierska A.; Benfenati E. 2003. Modeling toxicity by using supervised Kohonen neural network. *J. Chem. Inf. Comput. Sci.*, 43, pp. 485-492.

[132] Vracko, M. 1997. A study of Structure-Carcinogenic Potency Relationship with Artificial Neural Networks. The Using of Descriptors Related to Geometrical and Electronic Structures. *J. Chem. Inf. Comput. Sci.*, *37*, pp. 1037-1043.

[133] Vracko, M.; Novi , M.; Zupan, J. 1999. Study of structure-activity relationship by a counterpropagation neural network. *Anal. Chim. Acta*, 384, pp. 319-332.

[134] Zupan, J.; Novi , M.; Li, X.; Gasteiger, J. 1994. Classification of multicomponent analytical data of olive oils using different neural networks. *Anal. Chim. Acta*, *292*, pp. 219-234.

[135] McFarland, J.W. 1970 On the Parabolic Relationship Between Drug Potency and Hydrophobicity. *J. Med. Chem.* 13, pp. 1192-1196.

[136] Könemann, H. 1981 Quantitative structure-activity relationships in fish toxicity studies Part 1: Relationship for 50 industrial pollutants. *Toxicology*, 19, 3, pp. 209-221.

[137] Cronin, M.T.D.; Schultz, T.W. 1997. Validation of Vibrio fisheri acute toxicity data: mechanism of action-based QSARs for non-polar narcotics and polar narcotic phenols. *Sci. Total Environ.*, 204, 1, pp. 75-88.

[138] Ramos, E.U.; Vermeer, C.; Vaes, W.H.J.; Hermens, J.L.M. 1998. Acute toxicity of polar narcotics to three aquatic species (Daphnia magna, poecilia reticulata and Lymnaea stagnalis) and its relation to hydrophobicity. *Chemosphere*, 37, 4, pp. 633-650.

[139] Öberg, T. 2004. A QSAR for Baseline Toxicity: Validation, Domain of Application, and Prediction. *Chem. Res. Toxicol.*, 17, 12, pp. 1630-1637.

[140] Lipnick. R.L. 1991. Outliers: their origin and use in the classification of molecular mechanisms of toxicity. *Sci. Total Environ.* 109-110, pp. 131-153.

[141] Hadamard, J. 1923 Lectureson the Cauchy Problem in Linear Partial Differential Equations. Yale University Press.

[142] Smiesko, M.; Benfenati, E. 2004. Predictive Models for Aquatic Toxicity of Aldehydes Designed for Various Model Chemistries. *J. Chem. Inf. Comput. Sci.*, 44, 3, pp. 976-984.

[143] Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. 2005. Review of methods for QSAR applicability domain estimation by the training set. *ATLA*. (in press).

[144] Eriksson, L.; Jaworska, J.; Worth, A.; Cronin, M.T.D.; McDowell, R.M.; Gramatica, P. 2003. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression- Based QSARs. *Environ. Health Persp.*, 111, pp. 1351–1375.

[145] Duin, R.P.W.; Juszczak, P.; Paclik, P.; Pekalska, E.; de Ridder, D.; Tax, D.M.J. 2004. PRTools4, A Matlab Toolbox for Pattern Recognition, Delft University of Technology.

[146] http://www.prtools.org/prtools.html

[147] Fix, E.; Hodges, J.L. 1951. Discriminatory analysis, non-parametric discrimination. Technical report, USAF School of Aviation Medicine, Randolf Field, Tex. Project 21-49-004, Rept. 4, Contract AF41(128)-31.

[148] Vapnik, V.N. 1998. Statistical Learning Theory, John Wiley & Sons: New York.

[149] Cristianini, N.; Shawe-Taylor, J. 2000. An introduction to Support Vector Macines; Cambridge University Press: Cambridge, UK.

[150] Bennet; K.P., Campbell, C. 2000. Support Vector Machines: Hype or Hallelujah? *SIGKDD Explorations*, 2, 2, pp. 1-13.

[151] Schölkopf B., Smola A. Learning with Kernels; MIT Press: Cambridge, MA, 2002.

[152] Shannon, C.E. 1948. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27, pp. 379-423.

8. TABLES AND FIGURES

9. LIST OF PUBLICATIONS

The work in this thesis has contributed to the following publications.

9.1 CHAPTERS IN BOOKS

- Lo Piparo, E.; Fratev, F.; Mazzatorta, P.; Smiesko, M.; Benfenati E. Toxicity in allelopathy: *in silico approach*. In Allelopathy: A physiological process with ecological implications, Kluwer. (In press).
- Lo Piparo, E.; Mazzatorta, P.; Benfenati, E. Computational methods to Study Allelochemicals Properties. In Methods of Cellular diagnostics and analytical biochemistry, Narwal. (In progress).

9.2 PEER REVIEWED PAPERS IN JOURNALS

- Mazzatorta, P.; Benfenati, E.; Neagu, C.-D.; Gini, G. 2002. The importance of scaling in data mining for toxicity prediction. *J. Chem. Inf. Comput. Sci.*, 42, pp 1250-1255.
- Mazzatorta, P.; Benfenati, E.; Neagu, C.-D.; Gini, G. 2003. Tuning neural and fuzzy-neural networks for toxicity modeling. *J. Chem. Inf. Comput. Sci.*, 43, pp. 513-518.
- Mazzatorta, P.; Vra ko, M.; Jezierska, A.; Benfenati, E. 2003. Modeling toxicity by using supervised Kohonen neural network. *J. Chem. Inf. Comput. Sci.*, 43, pp. 485-492.
- Mazzatorta, P.; Vra ko, M.; Benfenati, E. 2003. ANVAS: Artificial Neural Variables Adaptation System for descriptor selection, *J. Comput Aid. Mol. Des.*, 17, pp. 335-346.

- Mazzatorta, P.; Benfenati, E.; Lorenzini, P.; Vighi, M. (2004), A QSAR approach to evaluate pesticide toxicity: an overview of modern local classification techniques. *J. Chem. Inf. Comput. Sci.*, 44, pp. 105-112.
- Lo Piparo, E.; Fratev, F.; Lemke, F.; Mazzatorta, P.; Smiesko, M.; Fritz, J.; Benfenati, E. QSAR models for Daphnia Magna toxicity prediction of benzoxazinoids and their transformation products. Submitted to Journal of Agricultural and Food Chemistry.
- Mazzatorta, P.; Smiesko, M.; Lo Piparo, E.; Benfenati, E. QSAR model for predictiving pesticides aquatic toxicity. Submitted to Chemical Research in Toxicology.
- Lo Piparo, E.; Smiesko, M.; Mazzatorta, P.; Indeger, J.; Bluemel, S.; Benfenati, E. CoMFA models to evaluate benzoxazinoids toxicity for Falsonia candida. Submitted to Journal of Agricultural and Food Chemistry.
- Mazzatorta, P.; Casalegno, M.; Benefenati, E.; Maran, U.; Sild, S. A QSAR Model for Acute Toxicity on Fathead Minnow (Pimephales promelas) using GRID technology. (In progress).

9.3 CONTRIBUTIONS IN CONFERENCE PROCEEDINGS

- Neagu, C.-D.; Benfenati, E.; Gini, G.; Mazzatorta, P.; Roncaglioni, A. 2002. Neuro-Fuzzy Knowledge Representation for Toxicity Prediction of Organic Compounds, Proceedings of ECAI 2002, 15th European Conference on Artificial Intelligence, July 21-26 2002, Lyon, France, pp. 498-502.
- Benfenati, E.; Mazzatorta, P.; Neagu, C.-D.; Gini, G. 2002. Combining classifiers of pesticides toxicity through a neuro-fuzzy approach. Proceedings of 3rd International Workshop on Multiple Classifier Systems MCS2002, 24-26 June 2002, Lecture Notes in Computer Science, Springer

Verlag Berlin Heidelberg, LNCS 2364, F. Roli and J. Kittler Eds., ISBN 3-540-43818-1, pp. 293-303, Cagliari, Italy.

- Mazzatorta, P.; Vra ko, M.; Jezierska, A. 2002. Toxicity Map: application of counterpropagation neural network in the investigation of aquatic acute toxicity, Proceedings of SKD 2002, Slovenski Kemijski Dnevi, September 26-27 2002, Maribor, Slovene, pp. 329-332.
- Jezierska, A.; Vra ko, M.; Mazzatorta, P. 2002. Counter-Propagation Artificial Neural Network Study on Prediction of the toxicity of Organic Compounds towards Fathead Minnow fish, Proceedings of SKD 2002, Slovenski Kemijski Dnevi, September 26-27 2002, Maribor, Slovene, pp. 317-322.
- Vra ko, M.; Mazzatorta, P.; Novic, M.; Roncaglioni, A.; Basak, S.; Mills, D. 2003. Counter Propagation Neural Network as a Tool in QSAR Modelling. Proceedings of SETAC Europe 13th Annual Meeting, Hamburg, Germany, 27 April - 1 May 2003.
- Vra ko, M.; Jezierska, A.; Valkova, I.; Mazzatorta, P.; Benfenati, E.; Basak, S.; Mills, D. 2003. Self Organizing map and Counter Propagation Neural Network as a Tool in Structure-Property Modelling. Proceedings of Third Indo-US Workshop on mathematical Chemistry, Duluth, Minnesota, August 2-7 2003.
- McCourt, D.; Lopez, J.; Benfenati, E.; Mazzatorta, P.; Romberg, M.; Schuller, B.; Dubitzky, W. 2003. Towards an Intelligent Data Type for Toxicity, Proceedings of IC-AI 2003, The 2003 International Conference on Artificial Intelligence, June 23-26 2003, Las Vegas, Nevada, USA.
- Mazzatorta, P.; Benfenati, E.; Schuller, B.; Romberg, M.; McCourt, D.; Dubitzky, W.; Sild, S.; Karelson, M.; Papp, A.; Bágyi, I.; Darvas, F. 2004.

OpenMolGRID: Molecular Science and Engineering in a Grid Context, Proceedings of PDPTA 2004, The 2004 International Conference on Parallel and Distributed Processing Techniques and Applications, Vol. II, 2004, pp. 775-779.

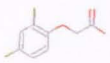
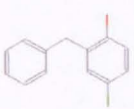
- Mazzatorta, P.; Lemke, F.; Benfenati, E. 2004. Application of self-organizing data mining for prediction of acute oraltoxicity to Bobwhite quail (Colinus Virginianus), The 11th International Workshop on Quantitative Structure-Activity Relationships in the Human Health and Environmental Sciences (QSAR 2004), May 9-13 2004, Liverpool, England.
- Mazzatorta, P.; Benfenati, E. 2004. Backpropagation neural network model for the prediction of soil sorption potential, The 11th International Workshop on Quantitative Structure-Activity Relationships in the Human Health and Environmental Sciences (QSAR 2004), May 9-13 2004, Liverpool, England.
- Vra ko, M.; Szymoszek, A.; Barbieri, P.; Jezierska, A.; Mazzatorta, P.; Valkova, I.; Bandelj, V. 2004. Self Organizing Maps (SOM) and Counter Propagation Neural Networks (CP NN) in Structure-Property Relationship Studies, The 11th International Workshop on Quantitative Structure-Activity Relationships in the Human Health and Environmental Sciences (QSAR 2004), May 9-13 2004, Liverpool, England.
- Benfenati, E.; Mazzatorta, P.; Lemke, F.; Gini, G.; Pintore, M.; Chrétien, J.; Chaudhry, Q.; Toropov, A. SOFT COMPUTING TECHNIQUES FOR TOXICITY PREDICTIONS: THE EXPERIENCES WITHIN THE PROJECT DEMETRA, submitted to Fuzzy logic, Soft Computing and Computational Intelligence Theories and Applications, World Congress (IFSA2005), July 28-31, 2005, Beijing China. (Accepted).

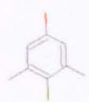
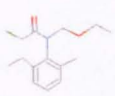
10. APPENDIX A: DATASET FOR ACUTE AQUATIC TOXICITY

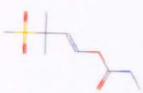
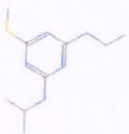
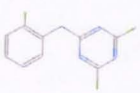
2D structure[12]	Name	CAS	Molecular Formula	[13]
	Pelargonic acid	112-05-0	C9 H18 O2	TR
	(Z)-11-Hexadecenal	53939-28-9	C16 H30 O	TR
	1,3-Dichloro-5,5-dimethylhydantoin(DCDMH)	118-52-5	C5 H6 N2 O2 Cl2	TR
	1,3-Dichloropropene	542-75-6	C3 H4 Cl2	TR
	1-Naphthylacetic acid	86-87-3	C12 H10 O2	TR

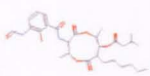
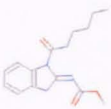
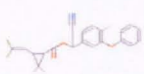
[12] Hydrogen atoms are omitted for clarity.

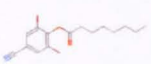
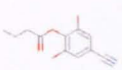
[13] TR: training set; TE: test set; V: validation set; E: eliminated.

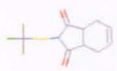
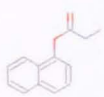
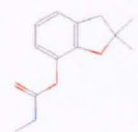
	2,4-D	94-75-7	C8 H6 O3 Cl2	TR
	2,4-D 2-butoxymethylethyl ester	1320-18-9	C15 H20 O4 Cl2	TE
	2,4-D butoxyethyl ester	1929-73-3	C14 H18 O4 Cl2	TR
	2,4-D butyl ester	94-80-4	C12 H14 O3 Cl2	TR
	2-Ethylhexyl 2-(2,4-dichlorophenoxy)propionate	79270-78-3	C17 H24 O3 Cl2	TR
	Clorophene	120-32-1	C13 H11 O Cl	TR

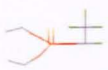
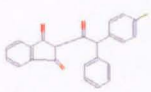
	3-Chloro-p-toluidine	7745-89-3	C7 H8 N Cl	TR
	4,4-dimethyloxazolidine	51200-87-4	C5 H11 N O	TR
	Kathon 930	64359-81-5	C11 H17 N O S Cl2	TR
	Chloroxylenol	88-04-0	C8 H9 O Cl	TR
	Acephate	30560-19-1	C4 H10 N O3 P S	TR
	Acetochlor	34256-82-1	C14 H20 N O2 Cl	TR

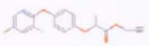
	Alachlor	15972-60-8	C14 H20 N O2 Cl	TR
	Aldicarb	116-06-3	C7 H14 N2 O2 S	TR
	Aldicarb sulfone	1646-88-4	C7 H14 N2 O4 S	TR
	Ametryne	834-12-8	C9 H17 N5 S	TR
	Amitraz	33089-61-1	C19 H23 N3	TR
	Anilazine	101-05-3	C9 H5 N4 Cl3	TR

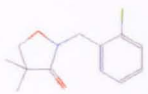
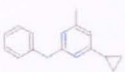
	Antimycin A	1397-94-0	C28 H40 N2 O9	TR
	Atrazine	1912-24-9	C8 H14 N5 Cl	TR
	Bendiocarb	22781-23-3	C11 H13 N O4	TE
	Benomyl	17804-35-2	C14 H18 N4 O3	TR
	Bensulide	741-58-2	C14 H24 N O4 P S3	TR
	Beta-Cyfluthrin	68359-37-5	C22 H18 N O3 F Cl2	TR

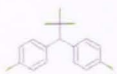
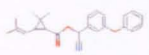
	Bis(trichloromethyl)sulfone	3064-70-8	C2 O2 S Cl6	TR
	Bromacil	314-40-9	C9 H13 N2 O2 Br	TE
	Bromoxynil	1689-84-5	C7 H3 N O Br2	V
	Bromoxynil octanoate	1689-99-2	C15 H17 N O2 Br2	TR
	Bromoxynil butyrate	3861-41-4	C11 H9 N O2 Br2	TR
	Bronopol	52-51-7	C3 H6 N O4 Br	TR

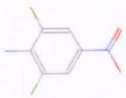
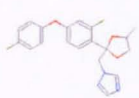
	Butralin	33629-47-9	C14 H21 N3 O4	TR
	Captan	133-06-2	C9 H8 N O2 S Cl3	TR
	Carbaril	63-25-2	C12 H11 N O2	TR
	Carbofuran	1563-66-2	C12 H15 N O3	TR
	Carboxin	5234-68-4	C12 H13 N O2 S	TR
	Chinomethionat	2439-01-2	C10 H6 N2 O S2	TR

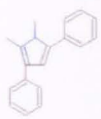
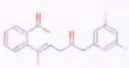
	Chlordecone	143-50-0	C10 O Cl10	TR
	Chloroethoxyfos	54593-83-8	C6 H11 O3 P S Cl4	TR
	Chlorhexidine	55-56-1	C22 H30 N10 Cl2	TR
	Chlorimuron ethyl	90982-32-4	C15 H15 N4 O6 S Cl	TR
	Chloroneb	2675-77-6	C8 H8 O2 Cl2	TR
	Chlorophacinone	3691-35-8	C23 H15 O3 Cl	TE

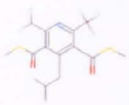
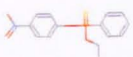
	Chloropicrin	76-06-2	C N O2 Cl3	TR
	Chlorpropham	101-21-3	C10 H12 N O2 Cl	TR
	Chlorpyrifos	2921-88-2	C9 H11 N O3 P S Cl3	TR
	Chlorpyrifos-methyl	5598-13-0	C7 H7 N O3 P S Cl3	TR
	Chlorthal-dimethyl	1861-32-1	C10 H6 O4 Cl4	TE
	Clodinafop-propargyl	105512-06-9	C17 H13 N O4 F Cl	V

	Clomazone	81777-89-1	C12 H14 N O2 Cl	TR
	Cycloate	1134-23-2	C11 H21 N O S	TR
	Cyhexatin	13121-70-5	C18 H34 O Sn	E
	Cymoxanil	57966-95-7	C7 H10 N4 O3	V
	Cyprodinil	121552-61-2	C14 H15 N3	TR
	Daminozide	1596-84-5	C6 H12 N2 O3	TR

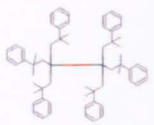
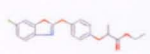
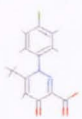
	DBNPA	10222-01-2	C3 H2 N2 O Br2	TR
	DDT	50-29-3	C14 H9 Cl5	TR
	Deltamethrin	52918-63-5	C22 H19 N O3 Br2	TE
	1,2-Dibromo-2,4-dicyanobutane	35691-65-7	C6 H6 N2 Br2	V
	Dichlobenil	1194-65-6	C7 H3 N Cl2	TR
	Dichlorprop	120-36-5	C9 H8 O3 Cl2	TR

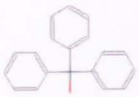
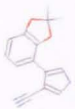
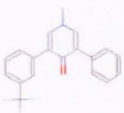
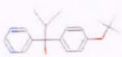
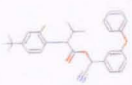
	Dichlorvos	62-73-7	C4 H7 O4 P Cl2	TR
	Dicloran	99-30-9	C6 H4 N2 O2 Cl2	TR
	Dicrotophos	141-66-2	C8 H16 N O5 P	TE
	Dieldrin	60-57-1	C12 H8 O Cl6	TE
	Dienochlor	2227-17-0	C10 Cl10	TE
	Difenoconazole	119446-68-3	C19 H17 N3 O3 Cl2	TR

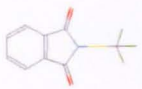
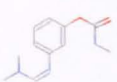
	Difenzoquat	49866-87-7	C17 H17 N2	TR
	Difethialone	104653-34-1	C31 H23 O2 S Br	TR
	Diflufenzopyr	109293-97-2	C15 H12 N4 O3 F2	TR
	Dimethenamid	87674-68-8	C12 H18 N O2 S Cl	TR
	Dimethoate	60-51-5	C5 H12 N O3 P S2	TR
	Diphacinone	82-66-6	C23 H16 O3	TE

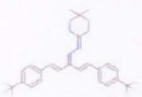
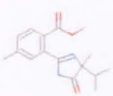
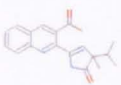
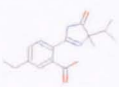
	Dithiopyr	97886-45-8	C15 H16 N O2 F5 S2	TR
	Diuron	330-54-1	C9 H10 N2 O Cl2	TR
	Dodine	2439-10-3	C13 H29 N3	TR
	Dowicil	4080-31-3	C9 H16 N4 Cl	TR
	Endrin	72-20-8	C12 H8 O Cl6	TR
	EPN	2104-64-5	C14 H14 N O4 P S	TR

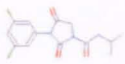
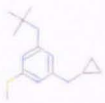
	EPTC	759-94-4	C9 H19 N O S	TR
	Ethalfluralin	55283-68-6	C13 H14 N3 O4 F3	TR
	Ethion	563-12-2	C9 H22 O4 P2 S4	TR
	Ethylene dichloride	107-06-2	C2 H4 Cl2	TE
	Farnesol	4602-84-0	C15 H26 O	TR
	Fenarimol	60168-88-9	C17 H12 N2 O Cl2	TR

	Fenbutatin oxide	13356-08-6	C60 H78 O Sn2	E
	Fenoxaprop-ethyl	66441-23-4	C18 H16 N O5 Cl	TE
	Fenoxycarb	79127-80-3	C17 H19 N O4	TR
	Fenpropathrin	39515-41-8	C22 H23 N O3	TR
	Fenridazone	68254-10-4	C12 H9 N2 O3 Cl	TE
	Fenthion	55-38-9	C10 H15 O3 P S2	TE

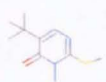
	Fentin hydroxide	76-87-9	C18 H16 O Sn	E
	Fludioxonil	131341-86-1	C12 H6 N2 O2 F2	TR
	Fluridone	59756-60-4	C19 H14 N O F3	TE
	Flurprimidol	56425-91-3	C15 H15 N2 O2 F3	TR
	Flutolanil	66332-96-5	C17 H16 N O2 F3	TE
	Fluvalinate	69409-94-5	C26 H22 N2 O3 F3 Cl	TR

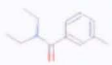
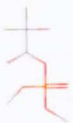
	Folpet	133-07-3	C9 H4 N O2 S Cl3	TR
	Fonofos	944-22-9	C10 H15 O P S2	TR
	Formetanate	22259-30-9	C11 H15 N3 O2	TR
	Fosamine	59682-52-9	C3 H8 N O4 P	TR
	gamma-HCH	58-89-9	C6 H6 Cl6	TR
	Glyphosate	1071-83-6	C3 H8 N O5 P	TR

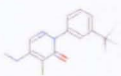
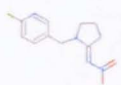
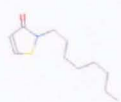
	Heptachlor	76-44-8	C10 H5 Cl7	TR
	(Z,E)-7,11-hexadecadienyl acetate	53042-79-8	C18 H32 O2	TE
	Hydramethylnon	67485-29-4	C25 H24 N4 F6	TR
	Imazamethabenz-methyl	81405-85-8	C16 H20 N2 O3	TR
	Imazaquin	81335-37-7	C17 H17 N3 O3	TR
	Imazethapyr	81335-77-5	C15 H19 N3 O3	TR

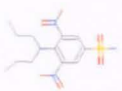
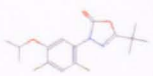
	Iprodione	36734-19-7	*C13 H13 N3 O3 Cl2*	*TR*
	Irgarol	28159-98-0	C11 H19 N5 S	TR
	Isofenphos	25311-71-1	C15 H24 N O4 P S	TR
	Isopropalin	33820-53-0	C15 H23 N3 O4	TR
	Limonene	138-86-3	C10 H16	TR
	Linuron	330-55-2	C9 H10 N2 O2 Cl2	TR

	perfluorooctane sulfonate	29457-72-5	C8 H O3 F17 S	TR
	Malathion	121-75-5	C10 H19 O6 P S2	TE
	Mecoprop	7085-19-0	C10 H11 O3 Cl	TR
	Mesotrione (AMBA)	104206-82-8	C14 H13 N O7 S	TE
	Metalaxyl	57837-19-1	C15 H21 N O4	TR
	Methidathion	950-37-8	C6 H11 N2 O4 P S3	TR

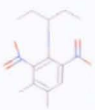
	Methomyl	16752-77-5	C5 H10 N2 O2 S	TR
	Methyl anthralinate	134-20-3	C8 H9 N O2	TR
	Methyl chloroform	71-55-6	C2 H3 Cl3	TR
	Methylene bis(thiocyanate)	6317-18-6	C3 H2 N2 S2	V
	Metolachlor-S-isomer	87392-12-9	C15 H22 N O2 Cl	TR
	Metribuzin	21087-64-9	C8 H14 N4 O S	TR

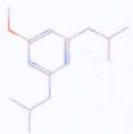
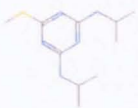
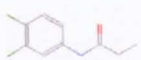
	MTI	82633-79-2	C7 H9 N O S	TR
	Diethyltoluamide	134-62-3	C12 H17 N O	TR
	Naled	300-76-5	C4 H7 O4 P Cl2 Br2	TR
	Napropamide	15299-99-7	C17 H21 N O2	TR
	Naptalam	132-66-1	C18 H13 N O3	TR
	Nerolidol	7212-44-4	C15 H26 O	TR

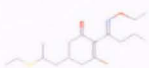
	Nitrapyrin	1929-82-4	C6 H3 N Cl4	TR
	N-methylneodecanamide	105726-67-8	C11 H23 N O	TR
	Norflurazon	27314-13-2	C12 H9 N3 O F3 Cl	TR
	Imidacloprid	105827-78-9	C9 H10 N5 O2 Cl	TR
	OBPA	58-36-6	C24 H16 O3 As2	E
	Octhilinone	26530-20-1	C11 H19 N O S	TR

	Oryzalin	19044-88-3	C12 H18 N4 O6 S	TR
	Oxadiazon	19666-30-9	C15 H18 N2 O3 Cl2	TR
	Oxamyl	23135-22-0	C7 H13 N3 O3 S	TR
	Oxazolidine E	7747-35-5	C7 H13 N O2	TR
	Oxydemeton-methyl	301-12-2	C6 H15 O4 P S2	TR
	Paclobutrazol	76738-62-0	C15 H20 N3 O Cl	TR

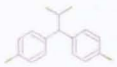
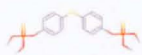
	Paradichlorobenzene	106-46-7	C6 H4 Cl2	TR
	Paranitrophenol	100-02-7	C6 H5 N O3	TR
	Parathion	56-38-2	C10 H14 N O5 P S	TR
	Parathion-methyl	298-00-0	C8 H10 N O5 P S	TR
	PCP	87-86-5	C6 H O Cl5	TR
	Pebulate	1114-71-2	C10 H21 N O S	TR

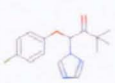
	Pendimethalin	40487-42-1	C13 H19 N3 O4	TR
	Phenmedipham	13684-63-4	C16 H16 N2 O4	TR
	Phorate	298-02-2	C7 H17 O2 P S3	TR
	Phosmet	732-11-6	C11 H12 N O4 P S2	TR
	Pindone	83-26-1	C14 H14 O3	TR
	Piperonyl butoxide	51-03-6	C19 H30 O5	TR

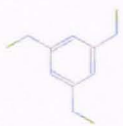
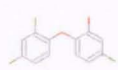
	Primisulfuron-methyl	86209-51-0	C15 H12 N4 O7 F4 S	TR
	Profenofos	41198-08-7	C11 H15 O3 P S Cl Br	TR
	Prometon	1610-18-0	C10 H19 N5 O	TE
	Prometryn	7287-19-6	C10 H19 N5 S	TR
	Propachlor	1918-16-7	C11 H14 N O Cl	TR
	Propanil	709-98-8	C9 H9 N O Cl2	TE

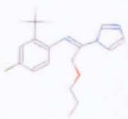
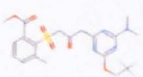
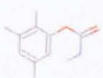
	Propargite	2312-35-8	C19 H26 O4 S	TE
	Propetamphos	31218-83-4	C10 H20 N O4 P S	TR
	Resmethrin	10453-86-8	C22 H26 O3	TR
	Rotenone	83-79-4	C23 H22 O6	TR
	Sethoxydim	74051-80-2	C17 H29 N O3 S	TE
	Siduron	1982-49-6	C14 H20 N2 O	TR

	Simazine	122-34-9	C7 H12 N5 Cl	TR
	2- Mercaptobenzothiazole	149-30-4	C7 H5 N S2	TR
	fluoroacetic acid	144-49-0	C2 H3 O2 F	TR
	Spinosad	131929-60-7	C41 H65 N O10	E
	Sulfotep	3689-24-5	C8 H20 O5 P2 S2	TE
	Tri-n-butyltin fluoride	1983-10-4	C12 H27 F Sn	E

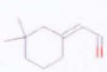
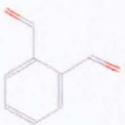
	TDE	72-54-8	C14 H10 Cl4	TR
	Tebupirimfos	96182-53-5	C13 H23 N2 O3 P S	TR
	Tebuthiuron	34014-18-1	C9 H16 N4 O S	TE
	Temephos	3383-96-8	C16 H20 O6 P2 S3	TR
	Terbacil	5902-51-2	C9 H13 N2 O2 Cl	TE
	Terbufos	13071-79-9	C9 H21 O2 P S3	TR

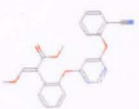
	Terbuthylazine	5915-41-3	C9 H16 N5 Cl	TR
	Thiazopyr	117718-60-2	C16 H17 N2 O2 F5 S	TR
	Thiobencarb	28249-77-6	C12 H16 N O S Cl	TR
	Thiram	137-26-8	C6 H12 N2 S4	TR
	Tralomethrin	66841-25-6	C22 H19 N O3 Br4	TR
	Triadimefon	43121-43-3	C14 H16 N3 O2 Cl	TR

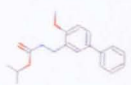
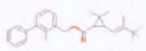
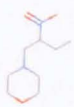
	Triadimenol	55219-65-3	C14 H18 N3 O2 Cl	TE
	Tri-allate	2303-17-5	C10 H16 N O S Cl3	TR
	Tribufos	78-48-8	C12 H27 O P S3	TR
	Trichloromelamine	7673-09-8	C3 H3 N6 Cl3	TR
	Triclopyr	55335-06-3	C7 H4 N O3 Cl3	TR
	Triclosan	3380-34-5	C12 H7 O2 Cl3	TR

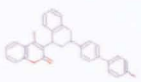
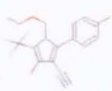
	Tridiphane	58138-08-2	C10 H7 O Cl5	TE
	Triflumidazole	68694-11-1	C15 H15 N3 O F3 Cl	TR
	Trifluralin	1582-09-8	C13 H16 N3 O4 F3	TR
	Triflusulfuron-methyl	126535-15-7	C17 H19 N6 O6 F3 S	TE
	Trimethacarb	2686-99-9	C11 H15 N O2	TR
	Tris nitro	126-11-4	C4 H9 N O5	TR

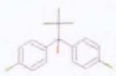
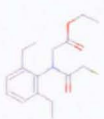
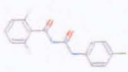
	Uniconazole	83657-17-4	C15 H18 N3 O Cl	TR
	Vernolate	1929-77-7	C10 H21 N O S	TR
	Tributyltin oxide	56-35-9	C24 H54 O Sn2	E
	Isopropanol	67-63-0	C3 H8 O	TR
	Propionic acid	79-09-4	C3 H6 O2	TR
	Tetrapropyl dithiopyrophosphate	3244-90-4	C12 H28 O5 P2 S2	TE

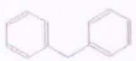
	1,3,5-Triethylhexahydro-s-triazine	7779-27-3	C9 H21 N3	TR
	(E)-(3,3-Dimethylcyclohexylidene)acetaldehyde	26532-25-2	C10 H16 O	TE
	3,5-dimethylpyrazole-1-carbinol	85264-33-1	C6 H10 N2 O	TR
	1,2-Benzenedicarboxaldehyde	643-79-8	C8 H6 O2	TE
	2,4-D Isopropyl Ester	94-11-1	C11 H12 O3 Cl2	TR
	2-Hydroxyethyl octyl sulfide	3547-33-9	C10 H22 O S	TR

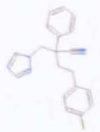
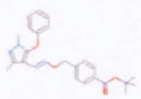
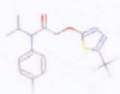
	3-Iodo-2-propynyl butylcarbamate	55406-53-6	C8 H12 N O2 I	TR
	4,5-Dichloro-1,2-dithiol-3-one	1192-52-5	C3 O S2 Cl2	TR
	Acibenzolar-s-methyl	135158-54-2	C8 H6 N2 O S2	TE
	Azoxystrobin	131860-33-8	C22 H17 N3 O5	TR
	Benfluralin	1861-40-1	C13 H16 N3 O4 F3	TR
	Bentazone	25057-89-0	C10 H12 N2 O3 S	TR

	Benzisothiazolin-3-one	2634-33-5	C7 H5 N O S	TR
	Beta cypermethrin	66841-24-5	C22 H19 N O3 Cl2	V
	Bifenazate	149877-41-8	C17 H20 N2 O3	TE
	Bifenthrin	82657-04-3	C23 H22 O2 F3 Cl	TR
	4-(2-nitrobutyl) morpholine	2224-44-4	*C8 H16 N2 O3*	TR
	2,2'-((1-methyl-1,3-propanediyl)bis(oxy))bis(4-methyldioxaborinane)	2665-13-6	C12 H24 B2 O6	E

	Brodifacoum	56073-10-0	C31 H23 O3 Br	TR
	beta-bromo-beta-nitrostyrene	7166-19-0	C8 H6 N O2 Br	TR
	Bromuconazole	116255-48-2	C13 H12 N3 O Cl2 Br	TR
	Carfentrazone-ethyl (F8246)	128639-02-1	C15 H14 N3 O3 F3 Cl2	TE
	Chlorfenapyr (Pirate)	122453-73-0	C15 H11 N2 O F3 Cl Br	TR
	Chlorflurenol methyl	2536-31-4	C15 H11 O3 Cl	TR

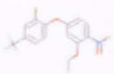
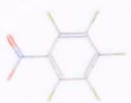
	Cimectacarb	95266-40-3	C13 H16 O5	TR
	Cyphenothrin	39515-40-7	C24 H25 N O3	TR
	DDAC	7173-51-5	C22 H48 N	TR
	Dicofol	115-32-2	C14 H9 O Cl5	TR
	Diethatyl ethyl	38727-55-8	C16 H22 N O3 Cl	TE
	Diflubenzuron	35367-38-5	C14 H9 N2 O2 F2 Cl	TR

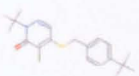
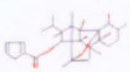
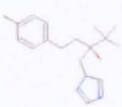
	Diiodomethyl p-tolyl sulfone	20018-09-1	C8 H8 O2 S I2	TR
	Dimethipin	55290-64-7	C6 H10 O4 S2	TR
	Diphenylamine	122-39-4	C12 H11 N	TR
	Dipropyl isocinchomeronate	136-45-8	C13 H17 N O4	TR
	DTEA	29873-30-1	C12 H27 N S	TR
	Ethoxyquin	91-53-2	C14 H19 N O	TR

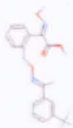
	Etridiazole	2593-15-9	C5 H5 N2 O S Cl3	TR
	Fenbuconazole	114369-43-6	C19 H17 N4 Cl	TR
	Fenhexamid	126833-17-8	C14 H17 N O2 Cl2	TR
	Fenpyroximate	134098-61-6	C24 H27 N3 O4	TR
	Fluazinam	79622-59-6	C13 H4 N4 O4 F6 Cl2	TE
	Flufenacet	142459-58-3	C14 H13 N3 O2 F4 S	TR

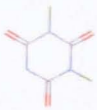
	Flumetralin	62924-70-3	C16 H12 N3 O4 F4 Cl	TE
	Flumiclorac pentyl	87546-18-7	C21 H23 N O5 F Cl	TR
	Flumioxazin (V-53482)	103361-09-7	C19 H15 N2 O4 F	TR
	Fluthiacet methyl	117337-19-6	C15 H15 N3 O3 F S2 Cl	TR
	Imazalil	35554-44-0	C14 H14 N2 O Cl2	TE
	Kresoxim methyl	143390-89-0	C18 H19 N O4	TR

	Lambda-Cyhalothrin	91465-08-6	C23 H19 N O3 F3 Cl	TR
	Methyl isothiocyanate	556-61-6	C2 H3 N S	V
	Methyl nonyl ketone	112-12-9	C11 H22 O	TR
	MGK 264	113-48-4	C17 H25 N O2	TR
	Naphthalene	91-20-3	C10 H8	TR
	Niclosamide	50-65-7	C13 H8 N2 O4 Cl2	TR

	O-Phenylphenol	90-43-7	C12 H10 O	TR
	Oxyfluorfen	42874-03-3	C15 H11 N O4 F3 Cl	TR
	Parachlorometacresol	59-50-7	C7 H7 O Cl	TR
	PCNB	82-68-8	C6 N O2 Cl5	TE
	Pirimicarb	23103-98-2	C11 H18 N4 O2	TR
	Pirimiphos methyl	29232-93-7	C11 H20 N3 O3 P S	TR

	Oleic acid	112-80-1	C18 H34 O2	TE
	Prallethrin	23031-36-9	C19 H24 O3	TE
	Pyridaben	96489-71-3	C19 H25 N2 O S Cl	TR
	Ryanodine	15662-33-6	C25 H35 N O9	V
	Strychnine	57-24-9	C21 H22 N2 O2	TR
	Tebuconazole	107534-96-3	C16 H22 N3 O Cl	TR

	Tefluthrin	79538-32-2	C17 H14 O2 F7 Cl	TE
	Thiodicarb	59669-26-0	C10 H18 N4 O4 S3	TE
	Butoxyethyl triclopyr	64700-56-7	C13 H16 N O4 Cl3	TR
	Trifloxystrobin	141517-21-7	C20 H19 N2 O4 F3	TR
	Zoxamide	156052-68-5	C14 H16 N O2 Cl3	TR
	Clopyralid	1702-17-6	C6 H3 N O2 Cl2	TR

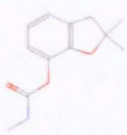
	Dichloroisocyanuric acid	2782-57-2	C3 H N3 O3 Cl2	TR

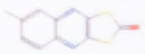
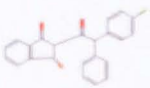
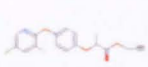
11. APPENDIX B: DATASET FOR AVIAN ORAL TOXICITY

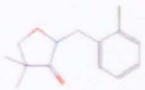
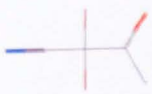
2D structure[14]	Name	CAS	Molecular Formula	[15]
	1,3-Dichloro-5,5-dimethylhydantoin(DCDMH)	118-52-5	C5 H6 N2 O2 Cl2	TR
	1,3-Dichloropropene	542-75-6	C3 H4 Cl2	TR
	Alachlor	15972-60-8	C14 H20 N O2 Cl	TR
	Aldicarb	116-06-3	C7 H14 N2 O2 S	TR
	Ametryne	834-12-8	C9 H17 N5 S	TR

[14] Hydrogen atoms are omitted for clarity.

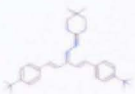
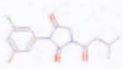
[15] TR: training set; V: validation set; E: eliminated.

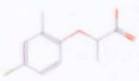
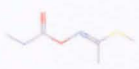
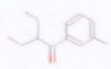
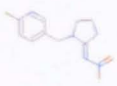
	Amitraz	33089-61-1	C19 H23 N3	TR
	Atrazine	1912-24-9	C8 H14 N5 Cl	TR
	Bendiocarb	22781-23-3	C11 H13 N O4	TR
	Bensulide	741-58-2	C14 H24 N O4 P S3	TE
	Oxydemeton-methyl	301-12-2	C6 H15 O4 P S2	TR
	Carbofuran	1563-66-2	C12 H15 N O3	TE

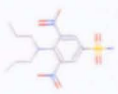
	Chinomethionat	2439-01-2	C10 H6 N2 O S2	TE
	Chloroethoxyfos	54593-83-8	C6 H11 O3 P S Cl4	TR
	Chlorhexidine	55-56-1	C22 H30 N10 Cl2	TR
	Chlorophacinone	3691-35-8	C23 H15 O3 Cl	TE
	Chlorpyrifos	2921-88-2	C9 H11 N O3 P S Cl3	TR
	Clodinafop-propargyl	105512-06-9	C17 H13 N O4 F Cl	TR

	Clomazone	81777-89-1	C12 H14 N O2 Cl	TR
	Cyhexatin	13121-70-5	C18 H34 O Sn	E
	DBNPA	10222-01-2	C3 H2 N2 O Br2	TE
	Dichlobenil	1194-65-6	C7 H3 N Cl2	TR
	Dichlorprop	120-36-5	C9 H8 O3 Cl2	TR
	Dichlorvos	62-73-7	C4 H7 O4 P Cl2	TR

	Dicloran	99-30-9	C6 H4 N2 O2 Cl2	TR
	Dienochlor	2227-17-0	C10 Cl10	TR
	Dimethenamid	87674-68-8	C12 H18 N O2 S Cl	TR
	Dodine	2439-10-3	C13 H29 N3	TR
	Dowicil	4080-31-3	C9 H16 N4 Cl	E
	Ethion	563-12-2	C9 H22 O4 P2 S4	TR

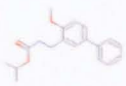
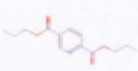
	Fenridazone	68254-10-4	C12 H9 N2 O3 Cl	TR
	Fenthion	55-38-9	C10 H15 O3 P S2	TR
	Formetanate	22259-30-9	C11 H15 N3 O2	TR
	Hydramethylnon	67485-29-4	C25 H24 N4 F6	TR
	Iprodione	36734-19-7	C13 H13 N3 O3 Cl2	TR
	Isofenphos	25311-71-1	C15 H24 N O4 P S	TR

	perfluorooctane sulfonate	29457-72-5	C8 H O3 F17 S	TR
	Mecoprop	7085-19-0	C10 H11 O3 Cl	TR
	Methomyl	16752-77-5	C5 H10 N2 O2 S	TR
	Diethyltoluamide	134-62-3	C12 H17 N O	TE
	Imidacloprid	105827-78-9	C9 H10 N5 O2 Cl	TR
	Octhilinone	26530-20-1	C11 H19 N O S	TR

	Oryzalin	19044-88-3	C12 H18 N4 O6 S	TR
	Paradichlorobenzene	106-46-7	C6 H4 Cl2	TE
	Paranitrophenol	100-02-7	C6 H5 N O3	TR
	Parathion-methyl	298-00-0	C8 H10 N O5 P S	TR
	PCP	87-86-5	C6 H O Cl5	TR
	Phorate	298-02-2	C7 H17 O2 P S3	TR

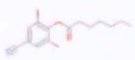
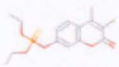
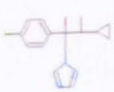
	Propachlor	1918-16-7	C11 H14 N O Cl	TE
	Propanil	709-98-8	C9 H9 N O Cl2	TR
	Tebupirimfos	96182-53-5	C13 H23 N2 O3 P S	TR
	Temephos	3383-96-8	C16 H20 O6 P2 S3	TR
	Terbufos	13071-79-9	C9 H21 O2 P S3	TR
	Thiazopyr	117718-60-2	C16 H17 N2 O2 F5 S	TR

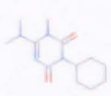
	Tribufos	78-48-8	C12 H27 O P S3	TR
	Triclosan	3380-34-5	C12 H7 O2 Cl3	TR
	Trimethacarb	2686-99-9	C11 H15 N O2	TR
	Uniconazole	83657-17-4	C15 H18 N3 O Cl	TR
	Dimethyl hydroxymethyl pyrazole	85264-33-1	C6 H10 N2 O	TR
	2,4-D Isopropyl Ester	94-11-1	C11 H12 O3 Cl2	TR

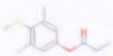
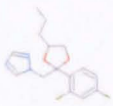
	3-Iodo-2-propynyl butylcarbamate	55406-53-6	C8 H12 N O2 I	TE
	4,5-Dichloro-1,2-dithiol-3-one	1192-52-5	C3 O S2 Cl2	TR
	Bentazone	25057-89-0	C10 H12 N2 O3 S	TE
	Bifenazate	149877-41-8	C17 H20 N2 O3	TR
	DDAC	7173-51-5	C22 H48 N	TR
	Dipropyl isocinchomeronate	136-45-8	C13 H17 N O4	TR

	Etridiazole	2593-15-9	C5 H5 N2 O S Cl3	TR
	Fluazinam	79622-59-6	C13 H4 N4 O4 F6 Cl2	TR
	Naphthalene	91-20-3	C10 H8	TR
	Parachlorometacresol	59-50-7	C7 H7 O Cl	TR
	Pirimiphos methyl	29232-93-7	C11 H20 N3 O3 P S	TR
	Prallethrin	23031-36-9	C19 H24 O3	TR

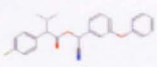
	Thiodicarb	59669-26-0	C10 H18 N4 O4 S3	TR
	Dichloroisocyanuric acid	2782-57-2	C3 H N3 O3 Cl2	TR
	2-(Hydroxymethylamino) ethanol	34375-28-5	C3 H9 N O2	TE
	Azinphos-methyl	86-50-0	C10 H12 N3 O3 P S2	TR
	Bromethalin	63333-35-7	C14 H7 N3 O4 F3 Br3	TR
	Bromo-3-chloro-5,5-dimethylhydantoin (BCDMH)	16079-88-2	C5 H6 N2 O2 Cl Br	TR

	Bromoxynil heptanoate	56634-95-8	C14 H15 N O2 Br2	TR
	Coumaphos	56-72-4	C14 H16 O5 P S Cl	TE
	Cyproconazole	94361-06-5	C15 H18 N3 O Cl	TR
	Diazinon	333-41-5	C12 H21 N2 O3 P S	TR
	Dicamba	1918-00-9	C8 H6 O3 Cl2	TR
	Diclofop-methyl	51338-27-3	C16 H14 O4 Cl2	TR

	Endosulfan	115-29-7	C9 H6 O3 S Cl6	TR
	Endothall	145-73-3	C8 H10 O5	TR
	Fenamiphos	22224-92-6	C13 H22 N O3 P S	TR
	Fenitrothion	122-14-5	C9 H12 N O5 P S	TR
	Fluchloralin	33245-39-5	C12 H13 N3 O4 F3 Cl	TR
	Hexazinone	51235-04-2	C12 H20 N4 O2	TE

	Hymexazol	10004-44-1	C4 H5 N O2	TR
	Methiocarb	2032-65-7	C11 H15 N O2 S	TE
	Methyl Bromide	74-83-9	C H3 Br	TR
	N6-Benzuladenine	1214-39-7	C12 H11 N5	TR
	Propiconazole	60207-90-1	C15 H17 N3 O2 Cl2	TR
	Sulfluramid	4151-50-2	C10 H6 N O2 F17 S	TR

	TCMTB	21564-17-0	C9 H6 N2 S3	TR
	4-Aminopyridine	504-24-5	C5 H6 N2	TE
	Chlorobenzilate	510-15-6	C16 H14 O3 Cl2	TR
	Chloroprop	101-10-0	C9 H9 O3 Cl	TE
	Cyromazine	66215-27-8	C6 H10 N6	TR
	Decyl isonomyl dimethyl ammonium	138698-36-9	C21 H46 N	TR

	Dimethoxane	828-00-2	C8 H14 O4	TR
	Dinoseb acid	88-85-7	C10 H12 N2 O5	TR
	Methanearsonate	144-21-8	C H5 O3 As	E
	Disulfoton	298-04-4	C8 H19 O2 P S3	TE
	Esfenvalerate	66230-04-4	C25 H22 N O3 Cl	TR
	Grotan	4719-04-4	C9 H21 N3 O3	TR

	MCPA Acid	94-74-6	C9 H9 O3 Cl	TE
	Mecoprop-P	16484-77-8	C10 H11 O3 Cl	TR
	Mefenoxam	70630-17-0	C15 H21 N O4	TR
	Methamidophos	10265-92-6	C2 H8 N O2 P S	TR
	Pyrithiobac	123342-93-8	C13 H11 N2 O4 S Cl	TR
	Dodecylbenzenesulfonate	27176-87-0	C18 H30 O3 S	TE

	Sulprofos	35400-43-2	C12 H19 O2 P S3	TR
	Trichlorfon	52-68-6	C4 H8 O4 P Cl3	TR
	Trichloro-s-triazinetrione	87-90-1	C3 N3 O3 Cl3	TR